OutRules: A Framework for Outlier Descriptions in Multiple Context Spaces

Emmanuel Müller, Fabian Keller, Sebastian Blanc, and Klemens Böhm

Karlsruhe Institute of Technology (KIT), Germany
{emmanuel.mueller,fabian.keller,klemens.boehm}@kit.edu,
sebastian.blanc@student.kit.edu

Abstract. Analyzing exceptional objects is an important mining task. It includes the identification of outliers but also the description of outlier properties in contrast to regular objects. However, existing *detection* approaches miss to provide important *descriptions* that allow human understanding of outlier reasons. In this work we present *OutRules*, a framework for outlier descriptions that enable an easy understanding of multiple outlier reasons in different contexts. We introduce outlier rules as a novel outlier description model. A rule illustrates the deviation of an outlier in contrast to its context that is considered to be normal. Our framework highlights the practical use of outlier rules and provides the basis for future development of outlier description models.

1 Open Challenges in Outlier Description

Outlier mining focuses on unexpected, rare, and suspicious objects in large data volumes [4]. Examples of outliers could be fraudulent activities in financial transaction records or unexpected patient behavior in health databases. Outlier mining has two aspects: (1) *identification* and (2) *description* of outliers. A multitude of approaches has been established for the former task (e.g., LOF [3] and more recent algorithms). They all focus on the quantification of outlierness, i.e., how strongly an object deviates from the residual data. Following this development of outlier detection algorithms there have been extensions of toolkits like *WEKA*, *RapidMiner*, and *R*, and stand-alone toolkits such as *ELKI* have been proposed. In all cases, only outlier detection algorithms have been implemented, and raw outlier results are visualized in different ways.

In contrast to this focus on the identification of outliers, approaches supporting outlier descriptions have been developed recently [6,1,2,7,9,10,5]. They aim at the description of the object's deviation, e.g. by selecting the deviating attributes for each individual outlier. These techniques assist humans in verifying the outlier characteristics. Without such outlier descriptions, humans are overwhelmed by outlier results that cannot be verified manually due to large and high dimensional databases. Humans might miss outlier reasons, especially if outliers are deviating w.r.t. multiple contexts. Therefore, humans depend on appropriate descriptions. This situation enforces the development of novel outlier description algorithms and their comparison in a unified framework.

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 828–832, 2012.

2 The *OutRules* Framework

With *OutRules*[1] we extend our outlier mining framework [8], which is based on the popular *WEKA* toolkit. *OutRules* extracts both regular and deviating attribute sets for each outlier and presents them as so-called outlier rules. We utilize the cognitive abilities of humans by allowing a comparison of the outlier object vs. its regular context. This comparison enables an easy understanding of the individual outlier characteristics. In a health-care example with attributes *age*, *height*, and *weight* (cf. Fig. 1), a description for the marked outlier could be "the outlier deviates w.r.t. (1) *height* and *weight* and (2) *height* and *age*". However, this first description provides the deviating attribute combinations only. In addition, we present groups of clustered objects (e.g., in attributes *weight* and *age*) as the regular contexts of the outlier. Overall, we present multiple contexts as regular neighborhoods from which the outlier is deviating. Reasoning is then enabled by manual comparison and exploration of these context spaces.

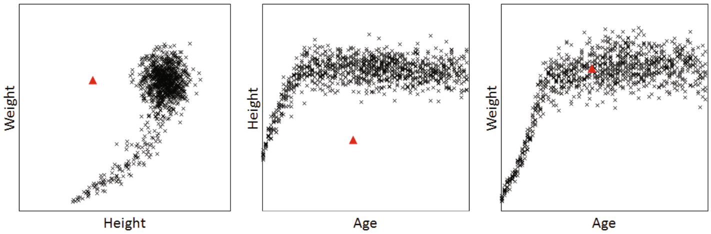


Fig. 1. Example of an outlier deviating w.r.t. multiple contexts

Outlier Rules as Basis for Outlier Descriptions

Our description model is based on the intuitive observation that each outlier deviates from other objects that are considered to be normal. Outlier rules accordingly represent these antagonistic properties of regularity on the one side and irregularity on the other side. As depicted in our example, there are multiple attribute combinations in which the object is an outlier, and there are multiple contexts in which it is regular. Several recent publications have observed this multiplicity of context spaces [1,7,9,10,5]. *OutRules* is the first framework that exploits these multiple context spaces for outlier rules. It illustrates the similarity among clustered objects and the deviation of the individual outlier. Therefore, it provides information about multiple contexts and highlights the differences to its local neighborhoods in these context spaces.

We consider each outlier individually and compute multiple outlier rules for each object. Each outlier rule is a set of attributes that show highly clustered objects on the one side, and on the other side, an extended set of attributes in which one of these objects is highly deviating. For instance in our previous example the outlier occurs under the attributes *age* and *height*. A first rule

[1] Project website: `http://www.ipd.kit.edu/~muellere/OutRules/`

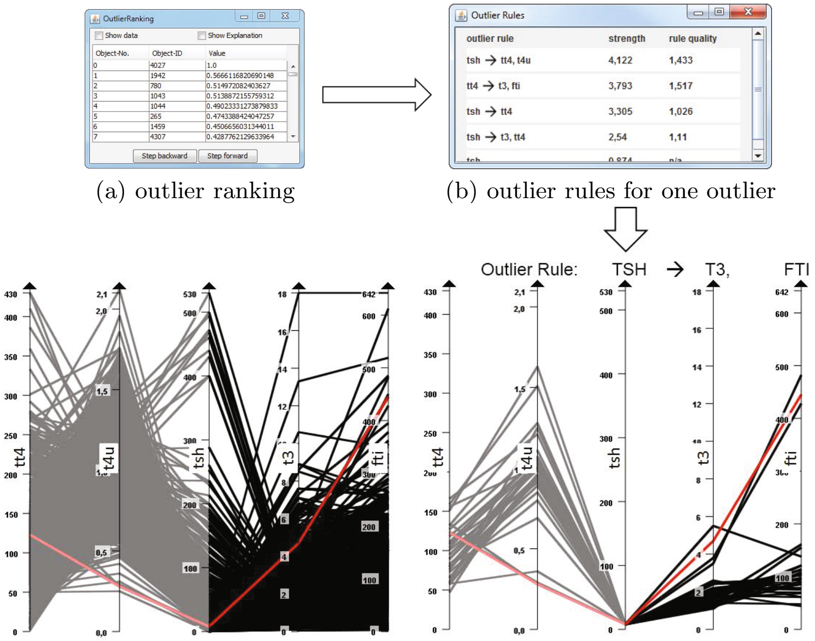
could be "The age is normal but the person is significantly too short". In this case the description might lead to the casual explanation that the represented person suffers from impaired growth. This outlier rule can be represented as $\{age\} \Rightarrow \{height\}$. Formally, an outlier rule is defined as follows:

Definition 1. ***Outlier Rule*** $A \Rightarrow B$

For an object o, the rule $A \Rightarrow B$ describes the *cluster membership* of o in attribute set $A \subseteq Attributes$ and the *deviating behavior* in $A \cup B \subseteq Attributes$.

The notion of *clustered* and *deviating* behavior can be instantiated by the underlying outlier score, e.g. by the notion of density in case of LOF [3].

We call A the *context* of o in which it shows regular behavior. As depicted in our example, there might be multiple reasons for an outlier deviation. Hence, our algorithm has to detect multiple contexts in which o is clustered. As the actual reason for an outlier is highly application-dependent, it is hard to make a binary decision of relevant and irrelevant rules. Therefore we output a ranking of all extracted rules. We rate each rule based on the data distribution in A and $A \cup B$. Based on the fact that an outlier rule represents the degree of regularity to other objects in the left hand side A and the degree of outlierness in the right hand side $A \cup B$, it is clear that the criteria have to quantify these two aspects. In our framework we have implemented criteria such as the *strength* of the outlier rule. It is defined as an instantiation of the well-established density-based outlier



Object-No.	Object-ID	Value
0	4027	1.0
1	1942	0.5666116820690148
2	780	0.514972082403627
3	1043	0.5138872155759312
4	1044	0.49023331273879833
5	265	0.4743388424047257
6	1459	0.4506656031344011
7	4307	0.4287762129633964

(a) outlier ranking

outlier rule	strength	rule quality
tsh → tt4, t4u	4,122	1,433
tt4 → t3, fti	3,793	1,517
tsh → tt4	3,305	1,026
tsh → t3, tt4	2,54	1,11
tsh	0,874	n/a

(b) outlier rules for one outlier

(c) parallel coordinates plots; left: no context; right: neighborhood in TSH

Fig. 2. One exemplary outlier from the Thyroid data set [UCI ML repository]

scoring [3]. Please note that the framework is open for any instantiation of quality criteria, e.g. for outlier rules in a specific application scenario.

Visualization of Outlier Rules

The visualization of outlier rules consists of three components. An overview of outliers is presented in the outlier ranking component (cf. Fig. 2(a)). Individual outliers can be chosen from this ranking for further exploration. The second component is a list of outlier rules sorted by the strength or other quality measures (cf. Fig. 2(b)). The last component is the visualization of individual outlier rules; each outlier rule can be explored in more detail by looking at the underlying data distribution. For example, we have implemented scatter plots, distribution statistics, density-distributions in individual attributes, and more enhanced visual representations such as well-established parallel coordinate plots (cf. Fig. 2(c)). As illustrated by the parallel coordinate plots for a real world example, *Thyroid Disease* from the UCI ML repository, the properties of the outlier rule and the nature of the outlier become clearer by the comparison with similar objects. If we consider all objects in the database in the left plot, we observe that the outlier is quite regular for all attributes from a global point of view. However, if one restricts the visualization to its local neighborhood in attribute TSH in the right plot there is a clear cluster containing the outlier, while attributes $T3$ and FTI show high deviation for the outlier from the local neighborhood. The clustering in TSH and the deviation in $\{T3, FTI\}$ indicate the correctness of this rule in a real world example.

Acknowledgments. This work has been supported by YIG "Outlier Mining in Heterogeneous Data Spaces" and RSA "Descriptive Outlier Mining" funded by KIT as part of the German Excellence Initiative; and by the German Research Foundation (DFG) within GRK 1194.

References

1. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: SIGMOD, pp. 37–46 (2001)
2. Angiulli, F., Fassetti, F., Palopoli, L.: Detecting outlying properties of exceptional objects. ACM Trans. Database Syst. 34(1), 1–62 (2009)
3. Breunig, M., Kriegel, H.-P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In: SIGMOD, pp. 93–104 (2000)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. 41(3) (2009)
5. Keller, F., Müller, E., Böhm, K.: HiCS: High contrast subspaces for density-based outlier ranking. In: ICDE, pp. 1037–1048 (2012)
6. Knorr, E.M., Ng, R.T.: Finding intensional knowledge of distance-based outliers. In: VLDB, pp. 211–222 (1999)
7. Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 831–838. Springer, Heidelberg (2009)

8. Müller, E., Schiffer, M., Gerwert, P., Hannen, M., Jansen, T., Seidl, T.: *SOREX*: Subspace Outlier Ranking Exploration Toolkit. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 607–610. Springer, Heidelberg (2010)
9. Müller, E., Schiffer, M., Seidl, T.: Statistical selection of relevant subspace projections for outlier ranking. In: ICDE, pp. 434–445 (2011)
10. Smets, K., Vreeken, J.: The odd one out: Identifying and characterising anomalies. In: SDM, pp. 804–815 (2011)