SOREX: Subspace Outlier Ranking Exploration Toolkit

Emmanuel Müller, Matthias Schiffer, Patrick Gerwert, Matthias Hannen, Timm Jansen, and Thomas Seidl

Data management and data exploration group
RWTH Aachen University, Germany
{mueller,mschiffer,gerwert,hannen,jansen,seidl}@cs.rwth-aachen.de
http://dme.rwth-aachen.de

Abstract. Outlier mining is an important data analysis task to distinguish exceptional outliers from regular objects. In recent research novel outlier ranking methods propose to focus on outliers hidden in subspace projections of the data. However, focusing only on the detection of outliers these approaches miss to provide reasons why an object should be considered as an outlier.

In this work, we propose a novel toolkit for exploration of subspace outlier rankings. To enable exploration of subspace outliers and to complete knowledge extraction we provide further descriptive information in addition to the pure detection of outliers. As wittinesses for the outlierness of an object, we provide information about the relevant projections describing the reasons for outlier properties. We provided *SOREX* as open source framework on our website[1] it is easily extensible and suitable for research and educational purposes in this emerging research area.

1 Challenges in Outlier Exploration

In general, the task of knowledge discovery in databases is twofold. On the one side data mining methods try to *detect* meaningful patterns, while on the other side knowledge is extracted out of the data by *providing descriptions* of these patterns. Especially, for the unsupervised outlier mining task, knowledge discovery does not end with the detection of the highly deviating objects. In applications like fraud detection, health surveillance, customer segmentation or sensor monitoring, one is interested in additional descriptions about the reasons why an object seems outlying. For example in health surveillance, a young patient might be considered as outlier due to high risk of dehydration (cf. o_1 in Figure 1). While dehydration is quite normal for elderly people, it is quite rare for young persons. By looking at a subset of measured attributes (*subspace projection*), one might detect this outlying patient showing high deviation from the residual patients in the attributes "age" and "skin humidity". While traditional outlier methods measure deviation using all attributes (*full data space*), subspace outlier ranking focuses on object deviation in subspaces.

[1] http://dme.rwth-aachen.de/OpenSubspace/SOREX

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 607–610, 2010.

However, not only the detection of such a high risk patient but also the underlying outlier properties are important. Providing information about the high deviation in age and skin humidity while showing normal measurements in all other attributes assists health professionals in verifying this automatically detected outlier. Thus, an obvious aim for outlier detection methods is to provide additional information about outlier properties such as the relevant attributes and to which extend a deviation from the regular objects can be observed. In Figure 1, we depict two outliers in three possible subspace projections for our toy example. As illustrated the hidden outliers show up only in specific projections while are hidden in other projections. For each outlier these specific projections can be seen as witnesses for its outlier properties.

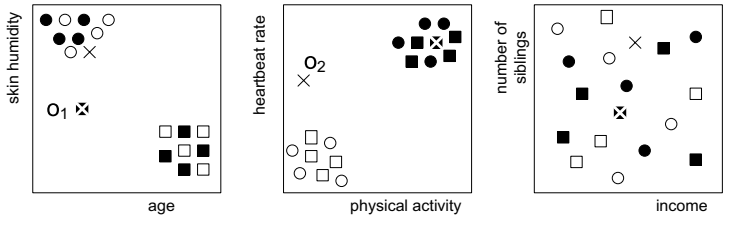


Fig. 1. Toy example (health surveillance): outliers hidden in subspaces

2 *SOREX*

Our novel SOREX toolkit provides such additional descriptions for each object. With SOREX we provide a repository of subspace outlier ranking algorithms [1–4], extending the popular WEKA framework. In addition, descriptive components provide the reasons why an object seems to be outlying. Overall, the main contributions of *SOREX* are:

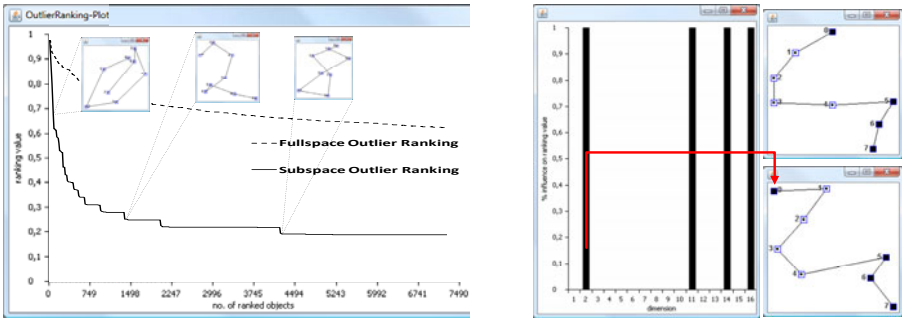
- Enhancing subspace outlier ranking by descriptive components.
- Open source toolkit for outlier exploration based on the WEKA framework

For our algorithm repository we include also some traditional outlier mining methods [5, 6]. In contrast to these full space methods, outlier detection in subspaces was first specified by [1], but without considering any additional outlier descriptions. Recent approaches have focused on outlier detection by considering projections of the data [2–4]. Their key idea is that outliers show high deviation from clustered objects in some subspaces. All of the proposed outlier ranking approaches have their focus on outlier *detection*, they provide only the ranking values as descriptive components. Thus, they are limited to providing a sorted list of most probable outliers, without giving explanations why an object seems to be an outlier. SOREX solves this drawback by additional descriptions.

Considering both mentioned contributions, *SOREX* is the first data mining framework for outlier mining that provides witnesses for the object's outlier properties. It enables the exploration of outliers and assists in reasoning of their

outlier properties. In contrast to existing data mining toolkits such as WEKA, RapidMiner, KNIME and RATTLE our toolkit focuses on the emerging research area of subspace outlier ranking not included in these frameworks. As part of our open source initiative "*OpenSubspace*" for subspace mining covering cluster detection, evaluation and visualization in our previous work [7–9], SOREX is a key component for exploration of outliers hidden in subspace projections.

Exploration based on descriptive components
In addition to the ranking values provided by the implemented outlier ranking algorithms SOREX provides descriptive components about relevant subspaces and the deviation in local neighborhoods. As post-processing to any subspace outlier ranking method SOREX can be used also for future algorithm enhancements in this emerging research area. In the following we describe the descriptive information in SOREX, illustrated also for an example in Figure 2.



(a) Outlier Ranking (b) Outlier Descriptor

Fig. 2. Descriptive outlier ranking on Pendigits data set from UCI ML repository

In general, the ranking value has to provide a clear distinction between outliers and regular objects. As depicted in Figure 2(a), subspace outlier rankings achieve high ranking values for the few objects supposed to be outliers. Thus, they can be clearly distinguished by the rapid decrease to the regular objects showing low ranking values. Using the full space or all possible subspaces for ranking results in the depicted bad ranking [2, 5, 6], where no clear distinction is possible. Their uniform decrease in this ranking lacks a clear support for the outlier detection. State-of-the-art ranking plots with additional visualization of ranked objects form the first descriptive components in SOREX. However, they do not yet provide reasons about ranking values, especially not about the relevant subspaces.

The relevant subspaces provide knowledge about the reasons why an object should be considered outlying. As depicted in Figure 2(b) for an outlying "four" in the pendigits database we derive a histogram describing the relative contribution of each attribute to the overall ranking value. One can observe which are the most deviating attributes for the considered object. This additional knowledge can be used to verify each outlier or even to provide its outlying properties.

In addition, one is not only interested in knowledge about the responsible attributes but also about the local neighborhoods of each outlier. Objects in these local neighborhoods appear as witnesses for the outlier properties as the outlier is highly deviating from this specific set of objects. By comparing the detected outlier with these objects one can extract the differences between a set of regular objects and one rare outlier as depicted for two version of the digit "four".

Demonstration of SOREX
The demo will illustrate the exploration of outlier ranking results for several data sets. It will allow conference attendees to explore the diverse subspace outlier ranking approaches implemented in the *SOREX* system, thus raising research interest in the area. Furthermore, the interfaces of our open source toolkit will facilitate the extension with further outlier mining algorithms and visual exploration paradigms by other researchers.

Our demonstration will raise interest for future work where the extracted knowledge about reasons for outlier properties could be used for interactive subspace outlier mining. Users may interact with the resulting descriptive outlier ranking by providing specific attributes which are supposed to be the reasons for outliers or focusing on specific sets of suspicious objects. This leads to in-depth analysis of the detected outliers.

Acknowledgments

This research was funded by the cluster of excellence on Ultra-high speed Mobile Information and Communication (UMIC) of the DFG (German Research Foundation grant EXC 89).

References

1. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: SIGMOD, pp. 37–46 (2001)
2. Lazarevic, A., Kumar, V.: Feature bagging for outlier detection. In: KDD, pp. 157–166 (2005)
3. Müller, E., Assent, I., Steinhausen, U., Seidl, T.: Outrank: Ranking outliers in high dimensional data. In: DBRank Workshop at ICDE, pp. 600–603 (2008)
4. Kriegel, H.P., Schubert, E., Zimek, A., Kröger, P.: Outlier detection in axis-parallel subspaces of high dimensional data. In: PAKDD, pp. 831–838 (2009)
5. Breunig, M., Kriegel, H.P., Ng, R., Sander, J.: LOF: identifying density-based local outliers. In: SIGMOD, pp. 93–104 (2000)
6. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: KDD, pp. 444–452 (2008)
7. Müller, E., Assent, I., Krieger, R., Jansen, T., Seidl, T.: Morpheus: Interactive exploration of subspace clustering. In: KDD, pp. 1089–1092 (2008)
8. Müller, E., Günnemann, S., Assent, I., Seidl, T.: Evaluating clustering in subspace projections of high dimensional data. In: VLDB, pp. 1270–1281 (2009)
9. Assent, I., Krieger, R., Müller, E., Seidl, T.: VISA: Visual subspace clustering analysis. ACM SIGKDD Explorations 9(2), 5–12 (2007)