OutRank: ranking outliers in high dimensional data

Emmanuel Müller [1], Ira Assent [2], Uwe Steinhausen [3], Thomas Seidl [4]

Data management and data exploration group
RWTH Aachen University, Germany
[1]mueller@cs.rwth-aachen.de
[2]assent@cs.rwth-aachen.de
[3]steinhausen@cs.rwth-aachen.de
[4]seidl@cs.rwth-aachen.de

***Abstract*— Outlier detection is an important data mining task for consistency checks, fraud detection, etc. Binary decision making on whether or not an object is an outlier is not appropriate in many applications and moreover hard to parametrize. Thus, recently, methods for outlier ranking have been proposed. Determining the degree of deviation, they do not require setting a decision boundary between outliers and the remaining data. High dimensional and heterogeneous (continuous and categorical attributes) data, however, pose a problem for most outlier ranking algorithms. In this work, we propose our OutRank approach for ranking outliers in heterogeneous high dimensional data. We introduce a consistent model for different attribute types. Our novel scoring functions transform the analyzed structure of the data to a meaningful ranking. Promising results in preliminary experiments show the potential for successful outlier ranking in high dimensional data.**

I. Introduction

Outlier detection is an important data mining task for managing today's huge amounts of application data. Applications include consistency checks of sensor network measurements, fraud detection in financial transactions, emergency detection in health surveillance and many more.

Outliers are objects that deviate from the rest of the data to a great extent. Distance-based [1] or cluster-based [2], [3] outlier mining algorithms suffer from difficulties in parametrization, as the extent of deviation is usually hard to quantify. This has led to outlier ranking based on their degree of deviation, e.g. as in the local outlier factor (LOF) approach [4] or in its extension to a top-n outlier detection [5]. While these approaches have been successful in low-dimensional data, high dimensional and heterogeneous data still pose a challenge to outlier detection. For high dimensional data (i.e. with very many attributes) as prevalent in many application databases, distances grow more and more alike due to an effect termed the "curse of dimensionality" [6]. As a consequence, both distance-based and clustering-based outlier detection methods fail to separate outliers from the remaining data. Global dimensionality reduction techniques like principal components analysis (PCA) [7] are not adequate in practical applications where the assumption of globally uniform relevance of attributes does not hold.

In clustering, the effect of locally varying relevance of attributes has led to the development of subspace clustering techniques. As clustering in the full space is no longer feasible, subspace clustering effectively detects locally relevant attributes (a lower dimensional subspace) for each cluster.

For high dimensional outlier ranking, we propose exploiting subspace clustering analysis. A major challenge lies in the fact that usually even outliers will be part of at least one of the exponentially many subspace clusters. Thus, the deviation has to be measured with respect to the prevailing subspace cluster patterns in the data. An additional challenge arises from heterogeneity, i.e. both continuous and categorical attributes, where existing approaches usually focus on just one type of attributes.

In this work we propose *OutRank* (*out*lier *rank*ing), an approach that is capable of handling heterogeneous high dimensional data. We introduce novel scoring functions to assess the deviation of objects from the rest of the data as determined by subspace clustering analysis. We therefore extend a recent subspace clustering model [8] to heterogeneous data in a consistent manner for both types of attributes. Preliminary experiments show that our algorithm outperforms LOADED [9], a link-based approach for heterogeneous data. We demonstrate future research potential toward a general framework for outlier ranking on arbitrary data types.

II. Outlier detection in subspaces

Outliers are objects that deviate from the overall data, e.g. as summarized in clustering results.

Challenge 1: **High dimensional data.**
Technique: subspace cluster analysis.

We propose using subspace clusters for outlier detection in high dimensional data where traditional clustering fails. The difficulty lies in meaningful ranking of outliers with respect to possibly overlapping clusters in arbitrary subspaces. As any object may belong to several clusters in one or more subspaces, we define novel scoring functions based on these subspace clusters (Sec. III).

In density-based (subspace) clustering, dense regions form clusters while objects in sparse regions are considered outliers. The density of an object o is measured via a density measure $\varphi(o)$ of objects in the ε-neighborhood of o. Objects in the neighborhood are "density-connected" and assigned to the same cluster.

Outlier detection requires comparable subspace clusters, i.e. the density measure φ_S has to be unbiased with respect to the dimensionality of the subspace S. This is achieved by normalizing with the expected density of the subspace dimensionality [8]:

Definition 1: **Unbiased density normalization**
For any density measure φ_S with expectation $E[\varphi_S]$,

$$\frac{\varphi_S}{E\left[\varphi_S\right]} \text{ is dimensionality unbiased.}$$

For continuous attributes, our previous work on dimensionality unbiased subspace clustering provides such an unbiased density measure [8]. Let $E_{cont}\left[\varphi_S\right]$ denote the expected density for a continuous valued subspace S. It is computed as the number of objects in the database DB multiplied by the volume ratio of the neighborhood in subspace S to the entire subspace S:

Definition 2: **Continuous normalization**

$$E_{cont}\left[\varphi_S\right] = |DB| \cdot \frac{vol(\varepsilon\text{-sphere}_S)}{vol(S)}$$

For computation details, please refer to [8].

Challenge 2: **Heterogeneous attributes.**
Technique: consistent density normalization.

For heterogeneous data, computation of the expected density requires taking categorical attributes into account. By definition, categorical data attributes have no extension, i.e. only discrete values occur. As a consequence, distance values are discrete as well and the notion of ε-sphere neighborhoods leads to discontinuous densities.

We propose a novel approach that unifies density assessment for categorical and continuous attributes. To ensure a consistent density measure, the expected density should be normalized for categorical attributes in the same manner as for continuous attributes. We achieve this consistency by treating categorical values not as discrete points, but as segments of the attribute value range. More precisely, the number of values v_i for each attribute dimension i of the categorical attributes is considered to be the value range extension in this attribute. The overall volume of a categorical subspace S_{cat} is then defined as the product of these ranges, yielding a rectangular overall volume $vol(S_{cat}) = \prod_{i \in S_{cat}} v_i$. The expected density of categorical attributes is then the number of objects in the database DB multiplied by the ratio of the segment volume by the volume of the subspace.

Definition 3: **Categorical normalization**

$$E_{cat}\left[\varphi_S\right] = |DB| \cdot \frac{vol(segment)}{vol(S)}$$

with $vol(segment) = 1$ as each segment corresponds to one discrete value. This view corresponds to a frequency count in the categorical attributes, and fits in nicely with our continuous attribute normalization in the sense that the overall expected density normalization $E\left[\varphi_S\right]$ is consistent for both types of attributes in subspace $S = S_{cont} \cup S_{cat}$:

Definition 4: **Heterogeneous normalization**

$$E\left[\varphi_S\right] = |DB| \cdot \frac{vol(\varepsilon\text{-sphere}_{S_{cont}})}{vol(S_{cont})} \cdot \frac{vol(segment)}{vol(S_{cat})}$$

Using this extended density measure definition for heterogeneous data, we determine subspace clusters (C, S) as maximal density-connected sets of objects C in a subspace S [8]. Subspace clusters according to this model are not redundant, i.e. clusters are not included in other clusters in higher dimensional subspaces.

Definition 5: **Subspace Clustering**
A subspace clustering w.r.t. to a density threshold F is a set $\{(C_1, S_1), \ldots, (C_n, S_n)\}$ of clusters C_i in subspaces S_i, i.e.

- C_i maximal, density-connected set of objects in S_i
- each object $o \in C_i$ is more dense than expected by at a least a factor F: $\varphi_{S_i}(o) \geq F \cdot E\left[\varphi_{S_i}\right]$
- C_i is not redundant in any higher dimensional subspace $S_j \supset S_i$, i.e. (C_i, S_j) not a subspace cluster

III. Outlier ranking

Using subspace clustering to analyze the structure of the data allows our approach to deal with high dimensional data. However, compared with traditional full space clustering algorithms like DBSCAN [2] there is no direct outlier output. Subspace clustering does not compute a partitioning of the data in several clusters and a group of outlier objects. Instead the result is a set of overlapping clusters which typically encloses all objects. We thus focus on defining scoring functions as a transformation based on subspace clusters. Scoring must reflect the deviation of objects such that a ranking of outliers can be computed by sorting objects in ascending order of their scores.

Challenge 3: **Outlier ranking on subspaces.**
Technique: scores for subspace cluster memberships.

In subspace clustering, objects are **typically in at least one** subspace cluster, because in low dimensional subspaces (e.g. one or two attributes) it is most probable to find similar objects. Additionally, objects are **often in more than one** subspace cluster, because of the large number of different subspaces, it is likely that each objects is similar to other objects in at least one subspace. Thus, we define outliers as objects that are found in abnormally few or low dimensional subspace clusters. Consequently, we develop novel scoring functions based on the result set of subspace clustering. For each object in the database these scoring functions assign positive score values for each object in each subspace cluster. The lower the score, the greater the deviation. Scoring functions should weight these clusters by their size with respect to both number of objects and number of attributes. This reflects the idea that larger subspace clusters in more attributes are stronger witnesses for an object's "normality".

In our first scoring function we incorporate cluster size $|C|$ and subspace dimensionality $|S|$ directly. The score is thus the weighted sum (by parameter α) of these two properties, normalized by maximal cluster size c_{max} and maximal dimensionality d_{max}:

Definition 6: **Size and dimensionality scoring**:

$$score_1(o) = \sum_{o \in (C,S)} \alpha \cdot \left(\frac{|C|}{c_{max}}\right) + (1-\alpha) \cdot \left(\frac{|S|}{d_{max}}\right)$$

An object o is assigned a score for each subspace cluster it belongs to, weighted by the size and dimensionality. Objects not in any subspace cluster score zero and objects in only small or very low dimensional subspace clusters score low, accurately reflecting their deviation from the prevailing patterns in the data.

Alternatively, in our second scoring function we use the measurements of density-based subspace clustering for outlier scoring. From the point of view of the density-based clustering paradigm, objects with high density are "normal". As discussed in the previous section, for subspace clustering this should be normalized by the expected density. Thus, the factor $\tilde{F}(o)$ by which an object o actually exceeds the expectation is an adequate weight for an object's score:

Definition 7: **Density expectation scoring**:

$$score_2(o) = \sum_{o \in (C,S)} \tilde{F}(o) = \sum_{o \in (C,S)} \frac{\varphi_S(o)}{E\left[\varphi_S\right]}$$

As before, objects not in any subspace clusters are assigned a value of zero, and objects that only barely exceed the expected density are given low scores.

For a preliminary evaluation of our approach we constructed a test data set[1] out of a real world database containing both categorical and continuous attributes. We randomly added 10 and 100 outliers, respectively, to assess the potential of the scoring functions for outlier ranking.

We measure recall and F1-measure values known e.g. from classification [10]. Recall is the ratio of outliers found in the ranking by the total number of outliers in the data. It indicates to which degree the rankings are successful in detecting the hidden outliers. The F1-measure is the harmonic mean of recall and precision, i.e. it also takes the false positives into account. For 10 outliers Figure 1 and Figure 2 show recall and F1-measure values vs. different ranking sizes for $score_1$, $score_2$ and the ranking used in the LOADED algorithm[2] [9]. Similarly, Figure 3 and Figure 4 show recall and F1-measure values for 100 outliers.

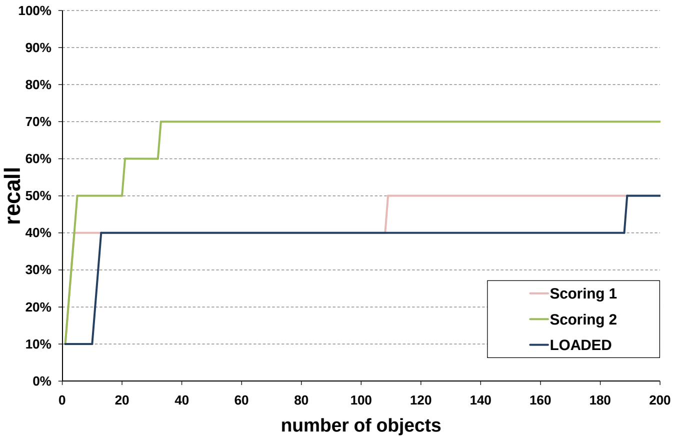

Fig. 1. Recall for 10 outliers

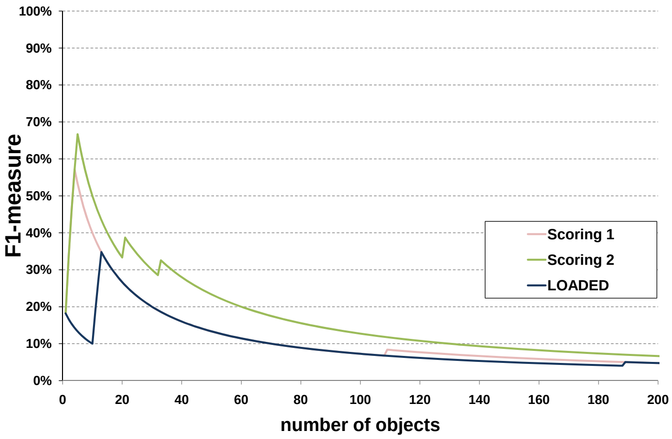

Fig. 2. F1-measure for 10 outliers

An ideal ranking should first find all hidden outliers and thus show both increasing recall and F1-measure until all outliers have been detected. Because we introduced outliers randomly, some of the introduced "outliers" may actually be consistent with the data distribution and not show up as outliers.

As we can see, our subspace clustering based approach yields promising results. It shows a faster increase in both recall and F1-measure, and in sum detects more outliers on the first 200 objects than the link-based competing algorithm. For the data set with only 10 outliers, $score_2$ shows a clearly better performance, while for 100 outliers, $score_1$ is slightly better at detecting outliers. This could be due to the fact that for more outliers, the factor by which the expected density of a subspace is exceeded better reflects the degree of deviation than the size and dimensionality of subspace clusters. We investigate this issue in our ongoing work to develop an overall optimal scoring function.

[1] data set containing 900 objects described by seven attributes. Earth quake monitoring database available at http://nsmp.wr.usgs.gov/data.html

[2] For $score_1$ we set $\alpha = 0.25$ by empirical evaluation.

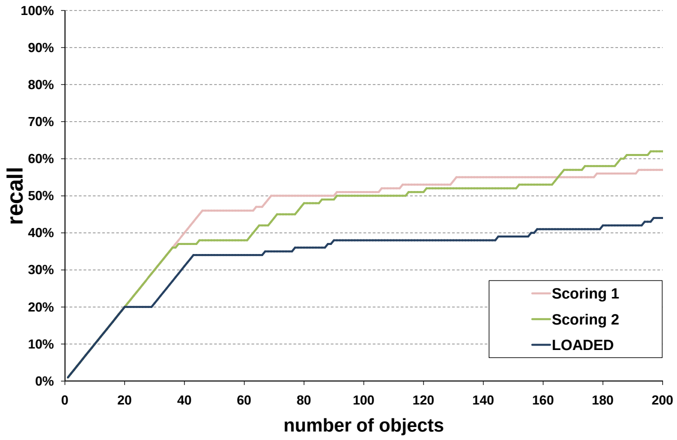

Fig. 3. Recall for 100 outliers

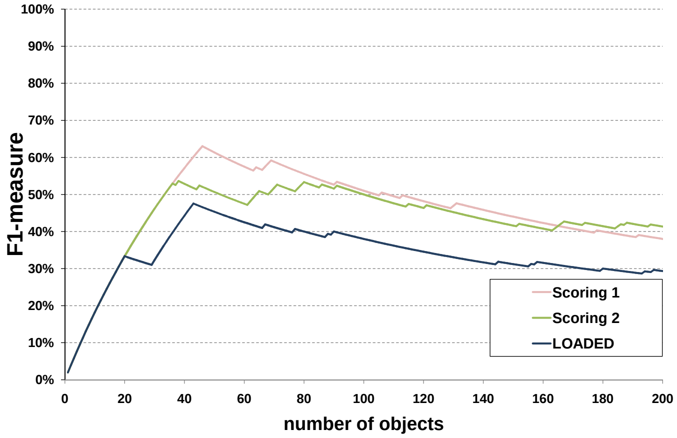

Fig. 4. F1-measure for 100 outliers

Compared with LOADED, which assigns link scores based on different schemes for categorical and continuous attributes, our consistent outlier measure for heterogeneous data shows superior performance. By generating distinct test sets with outliers in just categorical or just continuous attributes, we will investigate in future evaluations the differences to LOADED to further improve our scoring functions.

IV. Conclusion and future work

Ranking of outliers is a useful technique for the analysis of potentially deviating objects in the data. Starting from the most unusual objects with respect to patterns in the data, the ranking can be studied up to a user specified point. The ranking should thus reflect the degree of deviation. In this work, we have studied rankings that reflect outliers in the data.

As opposed to existing approaches, our OutRank approach handles heterogeneous data, i.e. both continuous and categorical attributes, of high dimensionality. Using subspace clusters as pattern representatives in high dimensional data, we have developed a novel normalization for detection of outliers in any subspace. The ensuing scoring functions reflect the main characteristics of the objects with respect to subspace clusters to ensure that outlier ranking reflects the degree of deviation.

Challenge 4: **Score analysis and combination.**
Technique: evaluation and interleaved top-k mining.

Our preliminary experiments show promising results for both scoring functions. Ongoing work deals with in-depth study of the performance of the two different scoring functions. Insights into their strengths and weaknesses could be combined into a framework that works for a broad range of application domains.

Moreover, we plan to interleave outlier ranking with the subspace cluster analysis. This allows for an immediate top-k outlier output as for low dimensional data in [5] before processing the whole ranking.

Challenge 5: **Sequential attributes.**
Technique: consistent normalization for sequences.

Additionally, OutRank will be extended to include sensor data as well. Taking the sequential nature of sensor data into account, an overall model that deals with continuous, categorical and sequential attributes allows application of our technique to any kind of data attributes.

Acknowledgment

This research was funded in part 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] E. Knorr, R. Ng, and V. Tucakov, “Distance-based outliers: algorithms and applications,” *VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[3] Z. He, X. Xu, and S. Deng, “Discovering cluster-based local outliers,” *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[4] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104.

[5] W. Jin, A. Tung, and J. Han, “Mining top-n local outliers in large databases,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 293–298, 2001.

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is nearest neighbors meaningful,” in *Proceedings of the International Conference on Database Theory*, 1999, pp. 217–235.

[7] I. Joliffe, *Principal Component Analysis*. Springer, New York, 1986.

[8] I. Assent, R. Krieger, E. Müller, and T. Seidl, “DUSC: Dimensionality unbiased subspace clustering,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2007, pp. 409–414.

[9] A. Ghoting, M. Otey, and S. Parthasarathy, “LOADED: Link-based outlier and anomaly detection in evolving data sets,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2004, pp. 387–390.

[10] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.