SubRank: Ranking local outliers in projections of high-dimensional spaces

Matthias Schiffer
matthias.schiffer@rwth-aachen.de

Data Mining and Data Exploration Group
Prof. Dr. Thomas Seidl
RWTH Aachen University, Germany

January 29, 2009

Abstract

Outlier mining has become an increasingly urgent issue in the KDD process, since it may be the case that finding exceptional events is more interesting than searching for common patterns. These outliers are most relevant to be found for instance in fraud detection processes. Unfortunately, existing approaches do not take into account that increasing dimensionality leads to a novel understanding of locality. Objects have to be investigated locally in different projections of the original space to overcome a crucial problem: The outlier property might be occluded by the shear number of dimensions. Being aware of this breach, in the course of my diploma thesis, I developed a novel, effective method, SubRank, to rank objects which are outliers only in some subspaces. Finally, it gives a concise explanation of the composition of the ranking itself.

Keywords: Data mining, high-dimensional data, outlier detection.

1 Introduction

Due to the fact that the amount of data being stored in databases increases rapidly, the need for methods in order to automatically investigate the data grows as well. *Knowledge discovery in databases* (KDD) has been defined as the process which deals with this task of retrieving non-trivial pieces of information from data. Many of its branches have been studied extensively, such as clustering or classification. However, outlier detection, the task of finding rare occasions in data sets, was none of them until the last decade. Till then, outliers were treated as unwanted noise and cleaned out.

In general, the assumption that outliers are unwanted noise does not hold true. Considering a simple e-commerce example makes this clear: Given a database of credit card transactions, it would be of special interest to find all the transactions which are suspicious to represent fraudulent behavior. Obviously, finding outliers is of highest priority in this scenario. To make things worse, because of the nature of high-dimensionality these true outliers are most probably occluded by the mere amount of attributes stored in a transaction.

In the last decade various methods have been proposed in order to identify outliers. Some of them are discussed representatively in Section 2. All methods have in common that they are not suitable in high-dimensional spaces. To get rid of this deficiency I worked out a new ranking method called SubRank. Experiments show that SubRank is capable of presenting objects, which are outliers in some subspaces but not necessarily in the whole space. Finally, this ranking is enriched with concise pieces of information, which give the user a guideline

for further investigation of the outlying objects in order to discover the application-dependent reasons for the outlier-property.

2 Related work

One of the very first methods is proposed in [KN98] and calculates the pairwise distances of objects to determine which of them are outliers. However, this methods makes a binary decision based on all attributes. As explained in the previous section, in many cases only a subset of the objects contributes to the property of being an outlier of a certain object which labels this method generally defective.

Due to this lack an alternative method called *LOF*, Local Outlier Factor, is proposed in [BKNS00] defining a new density-based notion of outlierness. Objects get a real-value assigned representing their degree of being an outlier and these values compose a ranking of the objects. Moreover, the value is assigned only based on calculations regarding the k-nearest neighbors of an object. The idea behind these calculations is the assumption that objects which are no outliers have relatively equal distances to their neighbors in comparison to the distances of the objects in the neighborhood of their neighbors respectively. Having this, the fundamental demand of locality regarding objects is fulfilled.

Unfortunately, LOF suffers from a major drawback. The k-nearest neighbors become less discriminant as the dimensionality increases. This effect is described in more detail in [BGRS99] and can be traced back to the well-known *curse of dimensionality*. Since both yet presented methods directly incorporate distances in full dimensionality, they are not applicable for finding outliers in high-dimensional spaces.

To overcome the scalability issue the authors of LOF present a further method in [KSZ08] called *ABOF*, Angle-Based Outlier Factor. Knowing that three points in a vector space define an angle, the basic concept of this approach can be summarized as follows: For every object calculate all angles by taking every pair of the remaining objects and compute the distance-weighted variance of these angles. A smaller variance shall indicate a higher probability that the considered object is an outlier than a higher one. Unfortunately, this approach does not consider the observation that objects do hardly group in high-dimensional spaces. Consequently, no outstanding objects can be found in the full space which would have a small range of measured angles. Finally, since simple distances are taken into account, this approach is doomed to suffer from the curse of dimensionality as well and hence fails to detect outliers in high-dimensional spaces.

A different concept named OutRank is introduced in [MASS08]. In general, outliers can also be recognized by analyzing a subspace clustering, because it appears to be a reasonable assumption that outliers are only present in very few or small subspace clusters. Based on this idea an arbitrary subspace clustering can be investigated to find such objects. However, the result of the analysis is highly dependent on the actual subspace clustering algorithm and the result is not based on all available pieces of information at the time of the clustering process, as stated in the paper. Making things even worse, OutRank does not profit from the insight of investigating subspaces and judges objects from a rather global view. As a consequence of this, OutRank's results can be very misleading or even suggest that an object is very unlikely to be an outlier. For instance, even if an object is an outstanding outlier in some projection, it might still show an average behavior in most other projections. Although such objects are highly relevant, the ranking position of such an object would be average. Obviously, this renders OutRank unreliable because the probability of actually observing such a situation grows as dimensionality increases.

Summing up all mentioned approaches, none of them is actually capable of finding outliers in high-dimensional spaces. Therefore, I present a novel idea of detecting outliers, which preseves the concepts of vicinity and ranking-values instead of binary decisions, that overcomes the problems arising from the high-dimensionality of the data.

3 Finding local subspace outliers

Though not applicable to high-dimensional data sets OutRank serves as the very foundation of my novel ranking method SubRank. OutRank analyzes a result of a subspace clustering and determines which objects are more likely to be outliers than others. So the idea that outliers can only be found by examining the subspaces of the original data set is somehow implicitly at hand, but is not used consistently.

Thus, I propose SubRank to build an adequate ranking. The key concept of this new ranking is that the basic choice whether an object o is a subspace outlier or not is determined by evaluating the *density* and its neighboring objects. Only if this density differs considerably from the others, the object will be treated as an outlier in this subspace. The density itself is computed by a *kernel density estimator*. Although an arbitrary kernel function could be chosen, the Epanechnikov kernel is an efficient and effective choice and provides the advantage that it assigns zero values to objects which do not fall into a certain area of influence controlled by a parameter h_d. With regard to the dimensionality bias, as dimensionality d increases this area has to grow as well by adjusting h_d to prohibit the underestimation of the density. Formally, the d-variate Epanechnikov kernel function is defined as:

$$K_e(x) = \begin{cases} \frac{d+2}{2c_d}(1 - \|x\|^2) & , \|x\| \leq 1 \\ 0 & , \text{else} \end{cases}$$

where c_d represents the volume of the d-dimensional unit sphere. Then the density f_d of o is defined, relative to its neighborhood $N_d(o) \subset DB$ depending on dimensionality d, as:

$$f_d(o) = \frac{1}{n(h_d)^d} \sum_{x \in N_d(o)} K_e\left(\frac{1}{h_d}\, d(o,x)\right) \quad \in [0,1]$$

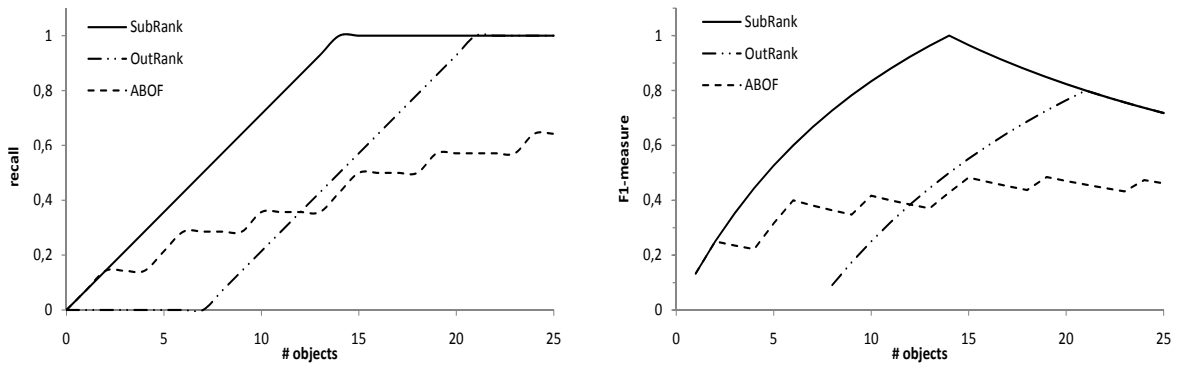
where $n = |N_d(o)|$ and $d(\bullet)$ represents an arbitrarily chosen distance function. Once the density of o is obtained, it is further scaled by the *degree* that it differs from the other densities to better reflect the overall degree of outlierness in this projection. In order to scale the density, it has to be tested whether o's density varies considerably which can be done by measuring the average of all densities in the neighborhood and then calculating the standard deviation of the densities. In statistics, a well-known assumption is that if a value deviates more than two times of the standard deviation around the mean, it should be a rare event. This provides a very simple and yet effective test to evaluate considerable difference. Thus, scaling is done by dividing the density by a factor $\text{degree}_d(o)$, where:

$$\text{degree}_d(o) = \frac{\text{avg} - f_d(o)}{2 \times \text{std\_dev}}$$

The whole procedure is repeated for every projection, where the object o appears to be a subspace outlier. Iff that is the case, then the scaled value is accumulated. This leads to the intuition that o is punished whenever a projection is found where o is an outlier. Consequently, o will get a higher rank, if it is found to be outlying in more projections. To achieve this goal entirely, the scaled density values are multiplied instead of summing them up. This guarantees that the ranking value will ever be higher than its smallest component. Hence, multiplication is the better choice to ensure that the order of the ranking is not blurred. Formally, a very abstract example for such undesired cases is given in the paragraph about OutRank in Section 2.

Finally, the resulting ranking will most probably deliver an overwhelming amount of subspace outliers, since the number of possible projections is exponential in the number of dimensions of the original data set. Confusing the user of the ranking is clearly an imminent risk. That makes it most urgent to deploy some hints as an fundamental ingredient of the ranking. Such hints include the projections in which the object was identified and the number of objects in the neighborhood, which are responsible for the outlier status.

4 Evaluation & Conclusion



(a) Recall (b) F1-measure

Figure 1: Quality of the experimentally obtained rankings

Proving the superiority of SubRank, I created a rather simple synthetic test scenario for preliminary experimental analysis. The data set contains 301 objects, each consisting of six continous dimensions, which form three subspace clusters. These subspace clusters are hidden twice in five and once in three dimensions. Additionally, 14 outliers are added so that the overall data set includes 315 objects. Concerning this data set, ABOF, OutRank and SubRank were run and the first 25 results built the foundation of the analysis of the result. To comprehensively analyze the ranking, I capture recall, the fraction of correctly ranked outliers so far, and F1-measure which is defined as the harmonic mean of precision and recall. Clearly, Figure 1(a) demonstrates that the curve of SubRank is saturated by reaching one as the first of the three approaches. This means that SubRank is the first to find all hidden outliers. Figure 1(b) shows the resulting graphs in terms of F1-measure. SubRank's curve progression is optimal since it discovers all 14 outliers at the top of the ranking and then ranks cluster objects leading to a decrease of values. The curve of OutRank has a rather similar progression but clearly a smaller peak which means that OutRank does rank some cluster objects (to be exact, seven) as outliers before the actual outliers follow. In contrast to that, ABOF is not even capable of finding all outliers, although two of them are ranked on the top positions. It ranks nine outliers spreaded over the top 25 positions which explains the curve progressing stepwise with skewed plateaus.

Summing up these results, SubRank is a very effective method to find outliers in the presence of high-dimensional data spaces. Incorporating the density directly into the ranking allows a much better reflection of the inherent data structure so that outliers can be found more reliable even in subspaces. According to the preliminary experiments, this approach is very promising in terms of efficiency, too. But due to space limitations these considerations are deferred to my diploma thesis as well as a detailed discussion about the descriptive components of the ranking.

References

[BGRS99] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 217–235, London, UK, 1999. Springer-Verlag.

[BKNS00] M. Breunig, H.P. Kriegel, R. Ng, and J. Sander. LOF: identifying density-based local outliers. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, New York, NY, USA, 2000. ACM.

[KN98] E. Knorr and R. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 392–403, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[KSZ08] H.P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452, New York, NY, USA, 2008. ACM.

[MASS08] E. Müller, I. Assent, U. Steinhausen, and T. Seidl. OutRank: ranking outliers in high dimensional data. In *Proc. 2nd International Workshop on Ranking in Databases (DBRank 2008) in conjunction with IEEE 24th International Conference on Data Engineering (ICDE 2008), Cancun, Mexico*, pages 600–603, 2008.