Outlier Detection and Ranking Based on Subspace Clustering

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

[1] RWTH Aachen University, Germany
{seidl, mueller}@cs.rwth-aachen.de
[2] Aalborg University, Denmark
ira@cs.aau.dk
[3] National Instruments Engineering GmbH, Germany
uwe.steinhausen@ni.com

Abstract. Detecting outliers is an important task for many applications including fraud detection or consistency validation in real world data. Particularly in the presence of uncertain data or imprecise data, similar objects regularly deviate in their attribute values. The notion of outliers has thus to be defined carefully. When considering outlier detection as a task which is complementary to clustering, binary decisions whether an object is regarded to be an outlier or not seem to be near at hand. For high-dimensional data, however, objects may belong to different clusters in different subspaces. More fine-grained concepts to define outliers are therefore demanded. By our new OutRank approach, we address outlier detection in heterogeneous high dimensional data and propose a novel scoring function that provides a consistent model for ranking outliers in the presence of different attribute types. Preliminary experiments demonstrate the potential for successful detection and reasonable ranking of outliers in high dimensional data sets.

1 Introduction

In many applications of scientific, engineering or business databases, the detection of outliers is an important task. Examples include the detection of rare events in large scale experiments, sensor failures in chemical process monitoring, fraud detection for credit card transactions, or alerting emergency situations in health care environments.

Conceptually, outlier detection can be regarded as a task which is complementary to clustering. Clustering aims at grouping similar objects to the same cluster whereas dissimilar objects tend to be assigned to different clusters. Objects which significantly deviate from the other objects may be identified as outliers and are not assigned to any cluster [1,2]. Rather than regarding these outliers to be just noise, outlier detection reports the objects to the application layer for further processing.

Several clustering algorithms have been proposed in the literature over the last decades. They are either based on a vector representation of the objects (e.g.,

Dagstuhl Seminar Proceedings 08421
Uncertainty Management in Information Systems
http://drops.dagstuhl.de/opus/volltexte/2009/1934

k-means) or they just rely on any distance function which indicates the dissimilarity of objects (e.g., k-medoid, DBSCAN, OPTICS). For high-dimensional data, however, many clustering methods fail to find clusters due to some effects known as "curse of dimensionality" [3]. Basically, the variance of distance values decreases with increasing dimensionality which means that similar and dissimilar objects cannot be discriminated well by looking at their distances over all dimensions.

In many situations, reduction of dimensionality helps, and Principal Components Analysis (PCA) [4] is a well known representative method. As a serious shortcoming, however, still many clusters do not show up in the reduced data space. As experiments and deeper analysis reveal, several clusters reside in different subspaces and, therefore, global reduction methods including PCA, Fourier or Wavelet transforms, do not support finding clusters which show up in different projections of the data space only.

Subspace clustering in general aims at identifying clusters in their individual projections of the data space. Various methods have been proposed in the literature to perform subspace clustering, projected clustering, or subspace search as some slightly different tasks are called in the field.

Whereas several concepts for cluster-based or distance-based outlier detection have been investigated, rarely any method for outlier detection takes the sketched subspace effects into account. As a particular problem, the original idea that either an object belongs to a cluster or the object is marked as an outlier is no longer valid as objects may be assigned to several clusters in different subspaces simultaneously. As an example, consider a database of persons which share various combinations of skills, experiences, hobbies, etc. Any person may be a member of the CS students cluster with data mining and Java skills and, in addition, may also belong to a cluster of soccer fans which is a hobby shared by students of other majors as well. Some persons may be assigned to several clusters but might be outliers with respect to the hobbies' domain whereas other persons are outliers with respect to all the domains. Thus, there are different choices for outlier models, and the notion needs to be adjusted in the world of high-dimensional data and the presence of subspace clusters. As our example demonstrates, "outlierness" can be observed in various degrees, and ranking factors for outliers seem to be an appropriate way to reflect the situation. In our new OutRank approach, we suggest ranking schemata for outliers based on their individual subspace situation [5].

2 Ranking of Outliers based on Subspace Clustering

The outcome of any subspace clustering algorithm is a set $\{(C_1, S_1), \ldots, (C_n, S_n)\}$ of subspace clusters (C_i, S_i) where C_i denotes the set of objects in the cluster and S_i indicates the respective dimensions of the cluster. In case of density-based subspace cluster definitions, a density measure $\varphi_S(o)$ is available for each element o from the data space. In order to rank objects with respect to their outlierness, we define an outlier score which applies to all objects o in the database.

Our first approach takes all the clusters (C, S) into account to which the object o belongs, and weighs their size $|C|$ and their dimensionality $|S|$ as follows:

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

In this case, an object o gets a high-valued score if it belongs to many big and high-dimensional subspace clusters. On the other hand, an object o which is a member of a few small and low-dimensional clusters gets scored low, and a strong outlier which does not belong to any cluster at all, is scored by a zero value. A competing approach takes the density estimation for the objects into account and defines the outlier score as follows:

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

Again, the scoring schema aggregates over all clusters (C, S) which contain the object o. The density estimator $\tilde{F}(o)$ reflects the density $\varphi_S(o)$ in a normalized way, i.e., divided by the expected density in the respective subspace. We thus follow the DUSC model of dimensionality-unbiased subspace clustering which has turned out to be an adequate model to cope with the mentioned observation of the curse of dimensionality [6]. In particular, the decreasing variance of distances with increasing dimensionality leads to smaller density values in high-dimensional spaces. In general, high-dimensional clusters tend to be quite sparser than low-dimensional clusters. According to DUSC, densities in different subspaces are normalized with respect to the expected density in those subspaces. DUSC thus is able to detect high-dimensional clusters while preventing from a flood of low-dimensional clusters by relating their density values to the expected density in the respective subspaces.

3 Evaluation

For our two outlier scoring schemata, we have performed some preliminary experiments on real data sets. These data are taken from an earth quake monitoring database[1] and contain 900 objects represented by seven attributes. We artificially added 100 outliers and compare our scoring schemata with LOADED [7], a link-based outlier scoring technique for heterogeneous data. We evaluate the retrieval quality by the standard F1-measure from information retrieval, computed as the harmonic mean of recall and precision [8]. In Figure 1 we compare the F1-values over a varying size of the database. The value of the weight alpha in score 1 is set to 0.25 which worked well also in other empirical evaluations. We observe in this and other experiments a good behavior of our new scoring schemata. Nevertheless, we are working on refinements and, in particular, on an automated weighting of the alpha weights.

[1] available from http://nsmp.wr.usgs.gov/data.html

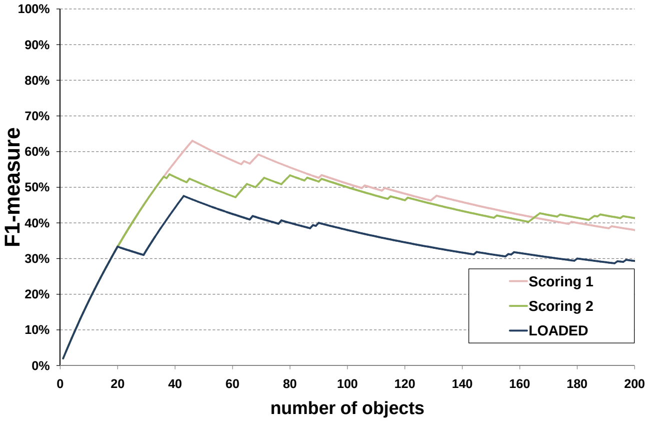

Fig. 1. F1-measure for 100 outliers

4 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. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: 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) 226–231
2. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recognition Letters **24** (2003) 1641–1650
3. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbors meaningful. In: Proceedings of the International Conference on Database Theory. (1999) 217–235
4. Joliffe, I.: Principal Component Analysis. Springer, New York (1986)
5. Müller, E., Assent, I., Steinhausen, U., Seidl, T.: Outrank: ranking outliers in high dimensional data. In: Proceedings of the 2nd International Workshop on Ranking in Databases (DBRank) in conjunction with IEEE 24th International Conference on Data Engineering (ICDE). (2008) 600–603
6. Assent, I., Krieger, R., Müller, E., Seidl, T.: DUSC: Dimensionality unbiased subspace clustering. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). (2007) 409–414
7. Ghoting, A., Otey, M., Parthasarathy, S.: LOADED: Link-based outlier and anomaly detection in evolving data sets. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). (2004) 387–390
8. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)