preprint - to appear at:
12th SIAM International Conference on Data Mining (SDM), Anaheim, CA, 2012

On Evaluation of Outlier Rankings and Outlier Scores

Erich Schubert Remigius Wojdanowski Arthur Zimek Hans-Peter Kriegel

Institut für Informatik, Ludwig-Maximilians-Universität München, Germany
http://www.dbs.ifi.lmu.de
{schube,wojdanowski,zimek,kriegel}@dbs.ifi.lmu.de

Abstract

Outlier detection research is currently focusing on the development of new methods and on improving the computation time for these methods. Evaluation however is rather heuristic, often considering just precision in the top k results or using the area under the ROC curve. These evaluation procedures do not allow for assessment of similarity between methods. Judging the similarity of or correlation between two rankings of outlier scores is an important question in itself but it is also an essential step towards meaningfully building outlier detection ensembles, where this aspect has been completely ignored so far. In this study, our generalized view of evaluation methods allows both to evaluate the performance of existing methods as well as to compare different methods w.r.t. their detection performance. Our new evaluation framework takes into consideration the class imbalance problem and offers new insights on similarity and redundancy of existing outlier detection methods. As a result, the design of effective ensemble methods for outlier detection is considerably enhanced.

1 Introduction

Outlier detection (i.e., the identification of points or objects that, for some reason, do not fit well to the remainder of a data set) is important for many application areas. Consequently, outlier detection is a quite active research area with many new methods proposed every year, based on different underlying methodologies like statistical reasoning [17], distances [2, 23, 36, 37, 44, 47], or densities [6, 8, 20, 24]. An open and widely ignored problem, though, is the proper and informative evaluation of the rankings of outlier scores provided by the different methods. Current outlier detection is often evaluated using *precision*@k, which is the true positive rate for the top k results in a data set that contains k outliers. This is a rather naïve way of evaluating the result due to the imbalanced nature of the problem: in a data set that contains 100 objects of which 2 are outliers, a method that ranks the true outliers on rank 3 and 4 will have a *precision*@2 of 0. Thus, occasionally, the precision values are shown for a range of k. This evaluation is sensible e.g. in information retrieval, where only the top 10 results of a search query are displayed, and the classification of the data set provides complete information on the relevancy. The task of unsupervised outlier detection however needs to accept that the "ground truth" may be incomplete and that real world data may include sensible outliers that are just not yet known or were considered uninteresting during labeling. Detecting such outliers however is overly punished due to class imbalance. A more advanced method of evaluation are receiver operating characteristic (ROC) curves, which plot the true positive rate against the false positive rate, thus inherently treating the class imbalance problem. To numerically compare ROC plots, one uses the area under this curve (AUC). However, this method also essentially loses the score information, i.e. whether the outlier score offers a reasonable contrast between outliers and inliers. Especially for ROC and focusing on classification tasks, this problem of losing the score information found attention recently [22]. Here, we try to take a wider perspective on evaluation of rankings, taking score information into account. We focus, though, on the application area of outlier detection.

An important motivation for our elaboration on an improved evaluation of outlier rankings and outlier scores is to support the design of ensemble methods for outlier detection. A key for building good ensembles is to use ensemble members that make uncorrelated errors. To meaningfully measure the correlation between the results of different methods for outlier detection is hence a step of paramount importance for building effective outlier ensembles. This has found no attention in previous attempts to ensemble outlier detection, presumably due to the lack of evaluation measures describing such correlation between methods.

We will discuss some reasoning about evaluation of (outlier) rankings presented in the literature in Section 2. In the following, we will elaborate on the relationship between common evaluation measures, and identify a general outlier score evaluation method (Section 3). We elaborate on requirements for building effective ensemble methods in Section 4. In the experiments (Section 5), we demonstrate the usefulness and benefits of the generalized evaluation measure by analyzing the similarity between results of different methods. Furthermore, we show how our evaluation measure is useful in enhancing an ensemble for outlier detection. Section 6 concludes the paper.

2 Related Work

Database research in ranking focuses on the scalability of the ranking retrieval algorithm (e.g. w.r.t. the top-k query problem [10, 12, 16, 33]). In the context of outlier detection, much work therefore aimed at making outlier detection more efficient in retrieving the top-k outliers, omitting the assessment of outlierness for those objects that, given the already received results, cannot rank among the top-k anymore. Examples are [2, 4, 19, 26, 36, 37, 44]. Usually, in these methods, the efficiency is evaluated while the quality of the ranking result is bound to be optimal w.r.t. the complete ranking (i.e., identical in the top k objects) as retrieved by the (less efficient) base method. Evaluation of rankings has not been a major issue in this research area. Recently, though, it was noted [32] that the evaluation of rankings of outlier scores merely based on ROC AUC values is unsatisfactory. Therefore, [32] introduced an "outlier ranking coefficient" of a ranking R as the ratio between the Spearman ranking coefficient between R and the best possible ranking and the Spearman ranking coefficient between the worst possible and the best possible ranking. The best ranking ranks all outliers before all inliers while the worst ranking ranks all inliers before all outliers. This does however not take into account that there are $k!$ possible "best" rankings for k outliers and hence does not tell anything about relationships of similarity or correlation between two different but possibly equally well rated rankings. The truly best outlier ranking should rank more obvious or clear outliers before less obvious outliers before those that could be outliers or inliers and so on. As reasoned in [13, 25], an optimally usable outlier score should provide an estimation of the probability of being an outlier. Ranking points according to their probability of being an outlier is then a truly meaningful ranking of objects (where ties, however, are possible). In the context of classification, ClasSi [18] is an abstraction and generalization of the reasoning in [32] that allows the use of multiple classes in the same ranking by giving a class similarity function. By choosing classes that resemble outliers and inliers, this can be applied to the evaluation of outlier scores as in the special case of [32], suffering from the same drawbacks. Neither of these two methods is suitable for comparing two different outlier ranking results meaningfully. Methods occasionally used for this task include Spearman's footrule, Spearman's ρ [41] and Kendall's τ [21]. These methods are designed in general for evaluating the similarity of two rankings. These values neither take the class imbalance problem into account, nor do they provide a good contrast for rankings where one expects the rankings to be strongly correlated. A weakness of these methods is that they solely evaluate the ranking, but not the underlying scoring. In addition, the common formula relies on a ranking with no ties, which is unrealistic in an outlier detection scenario where ties in particular with sure inliers are to be expected. Nevertheless, as time honored measures, both found a lot of attention in the information retrieval literature. In a very broad perspective, [38] discuss a couple of interpretations of Pearson's r. Top-k variants of Kendall's τ and Spearman's footrule are discussed in [11]. Melucci [31] discusses the observation that a single correlation coefficient (here: Kendall's τ) along with a test of significance gives limited information only. The use of Kolmogorov-Smirnov is proposed. A certain problem of Kendall's τ for the evaluation of rankings is that errors at any place are equally penalized. Usually, however, the top ranked items are most important. Shieh [39] introduces weights to penalize errors in different positions differently. An accuracy measure based on common elements in the top-k of both rankings, normalized by the number of objects of interest, is proposed in [45]. Yilmaz et al. [46] propose a mix of Kendall's τ and the average precision of relevant objects, thereby providing a probabilistic interpretation. Kendall's τ and Spearman's footrule are discussed and enriched in parallel in [27]. A more subtle problem of Kendall's τ has been pointed out e.g. in [7, 27, 40]: all swaps are treated as being statistically independent. In general, however, they are not. If two objects i, j share similar characteristics that are judged by one outlier model O_1 as outlierish, by another model O_2 they are not, swaps of both objects i, j with some other objects k, l in the two resulting outlier rankings are not independent. Instead of noticing the correlation between objects i, j, Kendall's τ is inclined to overestimate the correlation between the two outlier models.

While current correlation coefficients, like Blest's index [5], Shieh's weighted Kendall's τ [39], or τ_{AP} [46], do incorporate weights to give more importance to items being ranked at the top, they actually do not address the class imbalance problem. For outlier detection, balancing classes is necessary to restrict the influence of a large inlier class on the correlation coefficient value. Weights, if used to balance classes, should be related to the class size. The traditional measures use weights solely based on ranks [30], being not or at least not directly related to the class sizes. Furthermore, evaluation in information retrieval heads towards a different goal where precision or recall are not meaningful to individuals seeking only a few hits. However, for outlier detection, both recall and precision are important, as we wish to find as many outliers as possible while not miss-labeling too many inliers.

3 Generalization of Outlier Evaluation

3.1 Rank Similarity Measures A common formulation of Spearman's rank similarity is:

$$\rho = 1 - \frac{6\sum_i(\text{rank}_i(X) - \text{rank}_i(Y))^2}{n(n^2-1)}$$

This formula is only correct for a total ranking with no ties, that is each integer rank occurs exactly once in each ranking. Interestingly, it is a linear inversion of the squared Euclidean distance, applied to the rank position vectors $(\text{rank}_1, \ldots, \text{rank}_n)$. Another popular statistic, Spearmans footrule, can be seen as the Manhattan distance applied to the ranks. This leads to an intriguing parallel to the cost measure used in [25], which can be rewritten as a weighted Manhattan distance applied to vectors containing the "unified" outlier scores $(\text{unified}_1, \ldots, \text{unified}_n)$ to a target vector consisting of 0 for an inlier and 1 for an outlier. The full version of Spearman's rank similarity ρ however is defined as Pearson correlation applied to the ranks:

$$\rho = \frac{\text{Cov}(\text{rank}(X), \text{rank}(Y))}{\sqrt{\text{Var}(\text{rank}(X))}\sqrt{\text{Var}(\text{rank}(Y))}}$$

The Pearson correlation is a recurring theme in comparing rankings. Another popular correlation statistic, Kendall's τ, can be written as:

$$\begin{aligned}\tau &= \frac{1}{n(n-1)}\sum_{i\neq j}\text{sgn}(X_i - X_j)\text{sgn}(Y_i - Y_j)\\ &= E_{i\neq j}[\text{sgn}(X_i - X_j)\text{sgn}(Y_i - Y_j)]\end{aligned}$$

This however also is nothing but Pearson correlation, applied to the precedence relation: assign to each object combination $i \neq j$ a unique index to convert $C(X) := (c_{i,j} := \text{sgn}(X_i - X_j))$ into a vector field. Now consider the Pearson correlation of $C(X)$ and $C(Y)$. Note that $E[C(X)] = 0$ since $c_{i,j} + c_{j,i} = 0$. Without ties, we have $\text{Var}(C(X)) = 1$, since $c_{i,j}^2 = 1$. Then:

$$\begin{aligned}\rho_C &= \frac{\text{Cov}(C(X), C(Y))}{\sqrt{\text{Var}(C(X))}\sqrt{\text{Var}(C(Y))}} = \text{Cov}(C(X), C(Y))\\ &= E[(C(X) - E[C(X)])(C(Y) - E[C(Y)])]\\ &= E[C(X)C(Y)] = \tau\end{aligned}$$

There are many more such statistics. For example, Somers' D is the regression coefficient of $C(Y)$ with respect to $C(X)$ [34]. The ROC AUC statistic is equivalent to Somers' D [34]. Kruskal's γ is a modification of Kendall's τ to handle ties. ClasSi [18] is a generalization of Kendall's τ using weights derived from class similarity in particular for multiclass problems.

Essentially, all these measures of ranking similarity can be seen as a combination of two steps: (i) normalization and conversion; (ii) similarity computation. Popular choices are using the rank or order relation as normalization step and either Manhattan or Pearson correlation for the similarity computation. The popular formulas of the simplified Spearman's rank and Kendall's τ were developed as specializations of this pattern that offer computational benefits (by not having to compute variances) for integer ranks with no ties allowed.

3.2 Score Normalization As emphasized by the simplified formulas for Spearman's ρ and Kendall's τ, the common use of rank correlation is based on the assumption that the whole ranking is equally important and significant, whether the confused objects are ranked highly or not. An additional benefit is that, using ranks, arbitrary scores can be compared with each other, independent of their value range or meaning (as, e.g., in [28]).

Alas, this is not really appropriate for outlier detection: first, outlier detection is a rather imbalanced problem, where the key objects of interest are the rare class. Second, the ranking is significant only for the top objects, whereas the scores for the inliers usually do not vary much at all. In fact, the scores of inliers are often not even computed exactly, as e.g. in top-n approaches. Third, the scores themselves often convey a meaning and might even indicate that there are no outliers at all. Nevertheless, since algorithms often vary a lot in their value range and interpretation, it makes sense to perform some normalization on the score vectors. When comparing different methods that have different scales, some kind of normalization is unavoidable. Recently, [25] proposed a variety of normalization methods based on statistics and distribution fitting that can accommodate several outlier detection algorithms. These normalizations transform various outlier scores to a uniform value range of $[0\colon 1]$. We build upon those normalizations here to compare normalized outlier scores to the ground truth result, defined as 0 for true inliers and 1 for true outliers. We extend, however, the rather simple cost-based evaluation approach of [25] to a broader perspective and make it applicable to comparing multiple solutions with each other, instead of just measuring the cost with respect to the class labels.

Table 1: Summarized rank comparison measures

Simplified Spearman's ρ	
Normalization:	rank
Vector space:	score vectors
Measure:	normalized squared Euclidean
Full Spearman's ρ	
Normalization:	rank
Vector space:	score vectors
Measure:	Pearson correlation
Kendall's τ	
Normalization:	none
Vector space:	order relation
Measure:	Pearson correlation
ROC AUC	
Normalization:	none
Vector space:	order relation
Measure:	Regression coefficient
Unified Outlier [25]	
Normalization:	various normalizations proposed
Vector space:	score vectors
Measure:	weighted Manhattan distance
ClasSi [18]	
Normalization:	rank
Vector space:	class weighted order relation
Measure:	normalized Pearson

3.3 Outlier Scores as Vector Fields Usually, outlier scores have so far been only used to obtain an object ranking, expecting the outliers to come first. The "breadth first" ensemble method of [28] is an example of this interpretation, which merges the outlier scores of multiple algorithm runs to an ensemble by taking the top outlier each, then the second ranked objects each, etc., completely discarding the scores. However, a (numerical) outlier ranking can also be seen as a vector in a high dimensional space consisting of one dimension per object. By assigning a fixed index position to each object, one obtains a unique and complete representation of the outlier scores. The transformation is lossless, however it does not explicitly encode the rank information (unless we substitute the scores by their ranks, which can be seen as a special normalization). Given two outlier score vectors, the obvious way of comparing them is using an arbitrary distance function – e.g., Euclidean distance, Manhattan distance, or Pearson correlation distance. Due to the imbalanced nature of outlier scorings, we suggest however not to use the ordinary versions of these distances, but weighted variants.

3.4 Generalizing Existing Ranking Measures Seeing outlier score rankings as vector fields generalizes the traditional evaluation measures. Table 1 gives an overview of multiple rank comparison measures and how they fit into the general scheme discussed here. Some methods use the order relation instead of score vectors. Since it is not clear how to properly weight the order relation for an imbalanced problem and how to use the scores, we do not further explore this option in the remainder. Yet these can be seen as instances of the generalized view as well.

3.5 Choosing Appropriate Functions There is no "best" distance function, just as you cannot say whether Euclidean or Manhattan distance is "better" – this clearly depends on the problem you are trying to solve. But let us discuss a few rules of thumb for choosing appropriate functions for this framework.

Choosing a normalization and vector space: If the exact order of inliers and outliers within these groups is not important, particularly not among the inliers, a rank normalization is not appropriate. Instead, a score-based normalization such as the ones proposed in [25] is a good choice. If however there is little knowledge available about the actual scores and score distributions, a rank normalization can be useful. In our experiments, we will be only pursuing the first strategy, analyzing the actual scores.

Choosing a distance function: Since outlier detection is a highly imbalanced problem, applying this method to outlier scores requires a weighted measurement function. Secondly, the different distance functions can serve particular needs. Using Manhattan distance is good for estimating costs as it was intended and used in [25]. For optimization problems, the common procedure is to minimize squared errors, so a (weighted) Euclidean (or squared weighted Euclidean) is more appropriate. Finally, if we are interested in the correlation of scores and have little extra information about the score distribution, a correlation distance such as (weighted) Pearson correlation is the obvious choice, since it also includes some regularization of the scores by their mean and standard deviation.

In a supervised or semi-supervised context, the weights can be chosen according to the known outliers. In an unsupervised context, the weights can be estimated for example by considering the union of the top outliers of each method as true outliers.

3.6 Weighted Pearson Correlation Pearson correlation is an interesting candidate for our framework. However, the original method is unweighted, and thus will likely be strongly biased by the non-outliers. In order to apply it to an imbalanced problem, we employ a weighted covariance instead of the regular covariance. Let the weight assigned to object i be ω_i and $\Omega := \sum_i \omega_i$. The weighted mean for standardized weights is commonly defined as $E_\omega(X) := \frac{1}{\Omega} \sum_i \omega_i X_i$

Table 2: Artifical scorings of 4 positive and 4 negative examples, evaluated using typical measures

	A	B	B'	C	D	D'	E
+	1.0	1.0	1.0	1.0	1.0	1.0	1.0
+	1.0	1.0	1.0	1.0	1.0	1.0	0.0
+	0.9	0.9	0.9	0.55	0.1	0.1	1.0
+	0.3	0.2	0.1	0.55	0.01	0.0	0.0
−	0.1	0.1	0.2	0.45	0.0	0.01	1.0
−	0.0	0.1	0.1	0.45	0.0	0.0	0.0
−	0.0	0.0	0.0	0.0	0.0	0.0	1.0
−	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROC	0.000	0.000	0.188	0.000	0.000	0.312	1.000
smROC	0.038	0.056	0.193	0.225	0.019	0.330	1.000
Pearson	0.119	0.165	0.223	0.226	0.381	0.387	1.000
Sq. Eucl.	0.127	0.168	0.218	0.202	0.448	0.453	1.000
Manhattan	0.225	0.275	0.325	0.450	0.473	0.478	1.000

and the sample variance is then estimated using

$$\sigma^2_\omega(X) := \frac{1}{\Omega} \sum_i \omega_i (X_i - E_\omega(X))^2$$

Along this line, sample covariance is defined as

$$\mathrm{Cov}_\omega(X,Y) := \frac{1}{\Omega} \sum_i \omega_i \left(X_i - E_\omega(X)\right)\left(Y_i - E_\omega(Y)\right)$$

The weighted Pearson correlation then is:

$$\rho_\omega(X,Y) := \frac{\mathrm{Cov}_\omega(X,Y)}{\sigma_\omega(X)\sigma_\omega(Y)} \tag{3.1}$$

Weighted Pearson correlation, just like regular Pearson correlation, is in the range of $-1 \ldots +1$, and can be used as a dissimilarity function for example via $1 - \rho_\omega$.

3.7 Examples We study the effects of the proposed measure on a few toy examples, that also serve to highlight deficits in existing methods. In Table 2, we give a number of object scorings A to E. For simplicity we have chosen just 8 objects, half of them positive, half negative. We compute the scores of these rankings using a few typical methods. To make them easier comparable, we normalized the scores to a uniform range from 0 for "perfect" to 1 for "random". For ROC and smROC, we used $2(1 - AUC)$, for Pearson: $1 - \rho$. Squared Euclidean and Manhattan are linearly scaled.

Since ROC does not take the scores into account, it obviously is unable to differentiate A, B, C, and D, that all are perfect rankings. It does however consider the error from D to D' to be much more severe than B to B' while numerically it is the other way around. While the recent variant smROC [22] does use the scores, it behaves similarly here, because this setup actually triggers a non-continuity in its inversion step. It also considers D to be better than A. Pearson, squared Euclidean, and Manhattan essentially behave as expected in these cases, with no obvious benefits for either one. It can be seen that squared Euclidean, due to the use of squared values, is less punishing the "undecided" scoring of C close to 0.5 (which can be actually desirable in many situations).

3.8 Summary The general process of applying a score normalization, then computing a distance function on the score vectors is a rather obvious but so far underused general process in evaluating outlier scores. It can also be seen as a generalization of most existing measures, as shown in Table 1, which underlines the flexibility of this process and the many ways of optimizing it for particular situations. By using a weighted distance function, the imbalance problem can be handled. The wide choice of possible distance functions – including linear costs, quadratic errors, and correlation measures – makes the general principle usable in a wide range of domains, including computing the similarity (and correlation) of results, and quality measures of the scores, using classic mean squared error principles from statistics.

4 Enhancing Outlier Detection Ensembles

4.1 Background In classification, building ensembles of single classifiers to gain an improved effectiveness has a rich tradition and a sound theoretical background [9, 43]. Also in clustering, building ensemble clustering methods has found much interest over the years [15]. In the area of outlier detection, though much effort has been invested in design and implementation of advanced outlier detection algorithms, only some attempts to design superior combinations can be found in the literature [13, 25, 28, 35]. Let us reconsider the fundamental lessons learned w.r.t. ensemble methods in classification. The two basic conditions for an ensemble to improve over the contained base-classifiers are that the base classifiers themselves are (i) *accurate* (i.e., at least better than random) and (ii) *diverse* (i.e., making different errors on new instances). These two conditions are necessary and sufficient. If several individual classifiers were not diverse, then all of them will be wrong whenever one of them is wrong. This way, nothing is gained by combining them. On the other hand, if the errors made by the classifiers were uncorrelated, more individual classifiers may be correct while some individual classifiers are wrong. Therefore, a majority vote by an ensemble of these classifiers may be also correct. It is clear that each ensemble member should be at least somehow meaningful in order to get meaningful results out of their combination. Hence, a key for building good ensembles is to use ensemble members that make uncorrelated errors (if any). Though [28] introduced feature bagging (a common procedure in ensemble classification or ensemble clustering) to induce some diversity in several outlier detectors, the actually achieved decorrelation of detectors has not been evaluated. Overall, this requirement has not yet found theoretical attention in the few attempts to design outlier ensembles, and actually the means to do so have not been around so far. Instead, the four approaches known so far concentrated on methods for meaningfully combining the scores. They addressed the problem that scores delivered by different methods (or in different subspaces, where the scores are usually based on distances) usually vary strongly. For combination of such different scores, [28] proposed, first, a normalization by ranking (breadth-first traversal through the outlier rankings to combine), and, second, the cumulative sum of the different scores. The second decisively relies on the comparability of the retrieved scores. To enhance this comparability was the aim of the subsequently proposed approaches. Calibration approaches (sigmoid functions or mixture modeling) to fit outlier scores provided by different detectors into probability values have been used in [13]. They induce diversity by using different values for k of the kNN distance as an outlier score. In [35], the scores provided by a specific algorithm are centered around their mean and scaled by their standard deviation. Thus, all scores are in the same range, even if rather different algorithms are used for combination. Nevertheless, to induce diversity among different detectors, [35] primarily follow the feature bagging approach of [28]. Statistical reasoning was used in [25] to translate scores of different outlier detection methods into sort of outlier probabilities. There, the possibility of enhancement by combining different methods has been demonstrated, yet no measure of actual diversity or correlation between the used base algorithms has been applied.

4.2 Observations We will see in the experimental evaluation, for example, that building an ensemble from different LOF runs with different k (as in [13]) will not always yield a good ensemble since the results may be highly correlated. On the other hand, we will also demonstrate by means of our correlation analysis, that feature bagging indeed leads to rather uncorrelated results. Hence, it is in general reasonable to use feature bagging also as a heuristic for outlier detection ensembles. Alas, to enter a caveat, though the quality in combining outlier detectors from subspaces improves upon the single subspace outlier detectors, the quality can remain well below the quality of a single method using all features.

4.3 Greedy Ensemble Construction Based on analysis of correlation between outlier score rankings, we propose an unsupervised greedy ensemble construction approach, optimizing diversity: For a set I of individual outlier detectors, select the union of the top k data points of each instance in I, resulting in $K \ll k \cdot |I|$ different data points. Under the preliminary assumption that these were the true outliers of the data set, we build a target vector that has a 1 when the object is in these K objects and 0 otherwise. We compute the weights for our distance measure based on this working assumption using $\frac{1}{2K}$ and $\frac{1}{2(n-K)}$ as weights.

We initialize the ensemble with the detector $i \in I$ that has the highest weighted Pearson correlation to the target vector. Then we sort all remaining detectors $I \setminus i$ by the lowest correlation to the current ensemble, to maximize diversity. We test if we can improve our ensemble (increase the correlation with the target vector) by including the next method from the sorted list. If so, we update our ensemble with the additional method and reorder the list, otherwise we discard it. At each iteration, one method gets either included or discarded, and these decisions are not revised.

The motivation is that we have two factors to decide by: diversity and increase of correlation to the target vector (i.e., improved accuracy). Correlation increase obviously is the better hard factor – we do not want to include detectors that decrease our estimated performance – while diversity is more useful as a preference criterion. Also note that this process is entirely unsupervised (except for the initial choice of available detectors, obviously): The ensemble is greedily built from unsupervised detectors by just estimating their performance. Though this is a simple approach to merely demonstrate the applicability of a correlation measure in ensemble construction, the results are convincing and very competitive. Note that this greedy strategy is not possible with a classic *precision*@k or ROC evaluation, that cannot compare two methods for their diversity.

Clearly, using results with uncorrelated errors for combination can only be the second criterion since a certain level of accuracy in the ensemble members is essential – and the better the accuracy, the stronger the overall correlation. These two, however, are not as tightly connected as in classification or clustering, since we do not assess the correlation w.r.t. binary decisions but w.r.t. the resulting rankings, where quite different rankings can relate to equal performance in binary decision. This is exactly the gap in information that is not considered when using ROC analysis only. We demonstrate in the experiments how using this second criterion can improve the overall performance of an ensemble considerably.

5 Experiments

In the experiments, we use a weighted Pearson correlation distance as defined in Equation 3.1 for assessment of similarity of rankings of outlier scores, after normalizing the score vectors as suggested in [25].

5.1 Data The first data set is based on the "Amsterdam Library of Object Images" (ALOI) [14]. We extracted feature vectors of dimensionality 27 in HSB space. While the object classes contain about 110 each, we down-sampled objects to become rare arriving at a data set of 50000 objects, containing 1508 outliers. The "Wisconsin Breast Cancer" (WBC) and "Pen-Based Recognition of Handwritten Digits" (PenDigits) data sets are from the UCI machine learning repository [3]. To obtain outliers, again one of the classes (corresponding to malignant tumors in WBC and to '4' in PenDigits) was down-sampled to 10 (WBC) and 20 (PenDigits) instances. The Metabolic data set [29] measures the concentration of 43 metabolites in the blood of newborns. The control group (healthy) has $19,730$ instances, and there are 659 atypical records (illness). The KDDCup 1999 data set, while probably not resembling real network data and having irregularities, was also included in our experiments because of its size with $60,839$ records and 38 numerical dimensions (in the preprocessed version we use) and because it is commonly used in related work. For our experiments, we are using unsupervised methods only and it does not matter whether we are detecting actual network intrusions or just defects in the data set.

5.2 Similarity of Methods We compare a series of kNN-based outlier detection methods with each other: LOF [6], LDOF [47], LoOP [24], kNN [37], aggregated kNN [2], and ABOD [26], using a variety of parameter settings for each. (We omitted ABOD for the ALOI data set, where it did not finish in reasonable time.) We use the methods as implemented in ELKI [1].

Figure 1 gives a similarity matrix for these algorithms along with a ground truth (the result matching the data set labels) on the ALOI data. Dark signals high similarity (correlation) of the ranking results of two compared algorithms. For each algorithm, the parameter k is set to $\{5, 10, 15, 20, 25\}$ in turn. As can be estimated from the first column and row (correlation with ground truth), LoOP performed best (though far from good) in this experiment, LDOF did not find any outlier. Since the data are color histograms, we also used the histogram intersection distance [42] on a finer grained range of values for k. While the results are better overall, the similarity matrix looks virtually the same (see below, Figure 6(b)). Apparently, ALOI is not an easy target for all used algorithms, as the results are not highly correlated with the ground truth throughout.

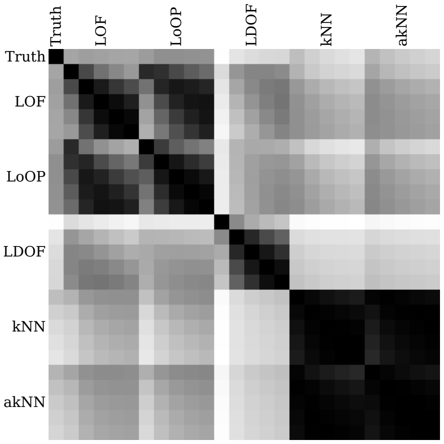


Figure 1: ALOI with Euclidean distance

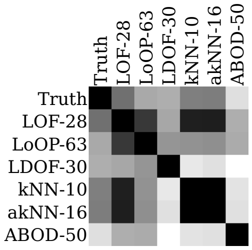
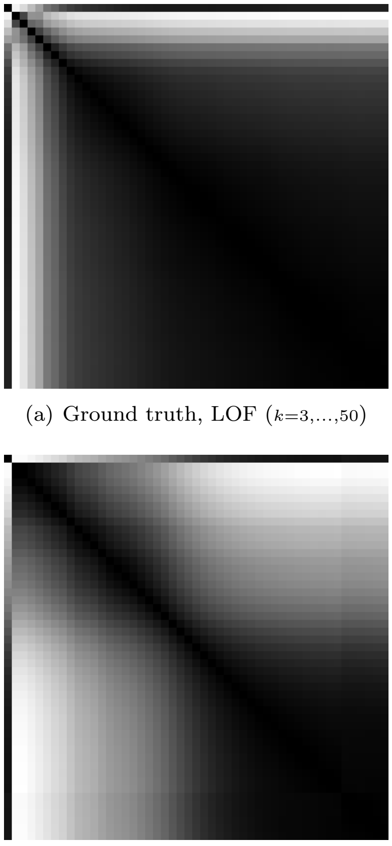


Figure 2: WBC with Euclidean distance

The WBC data are an easier target for all methods (here including ABOD: Figure 2). We selected the optimal parameter values here, using Euclidean distance, later we inspect the stability over varying parameter values (Section 5.3). Nevertheless, the methods yield outlier scores that are quite dissimilar w.r.t. each other. Note that all variants are actually performing well in terms of ROC AUC on WBC but the correlation assessment reveals disagreements in the rankings where, still, the top-ranked items (though ranked differently within the top-group) are mostly outliers.

The two methods based on kNN distances (kNN and aggregated kNN) are strongly correlated with each other but show the least similarity to all other methods. The aggregated version, though, is consistently a bit more similar to LOF, LoOP, and LDOF as the version solely based on a single distance. This observation makes sense, as the aggregation already introduces a dampening which is increased by the more sophisticated local methods.

5.3 Parameter Stability The parameter stability of methods can be inspected considering the diagonal



(a) Ground truth, LOF (k=3,...,50)

(b) Ground truth, LDOF (k=3,...,50)

Figure 3: WBC: Within method similarity over varying parameter k with weighted Manhattan distance

blocks of the matrices for the same method with different k. Considering, e.g., LOF and LoOP on ALOI (see Figure 6(b)), the rather smooth plot indicates a general stability over a broad range of k, i.e., the results of these runs with different parameter settings are highly correlated (except for rather high values of k vs. rather low values of k). These findings should discourage in general from building an ensemble consisting of LOF runs with different ks (irrespective of the additional issue of accuracy, which is also bad here).

A somewhat different behavior can be seen on the WBC data. Here, we inspect the stability of LOF and LDOF (Fig. 3) in detail. We see both methods improving with increasing $k = 3, \ldots, 50$, but with a different pace and different stability. LOF essentially needs a k chosen big enough to perform well. LDOF is less stable, a broad range of small k values retrieves results that are quite dissimilar to the (better) results with big k values. Other methods result in similar findings (not shown): LoOP is comparable to LOF in stability, the kNN model is even less stable than LDOF.

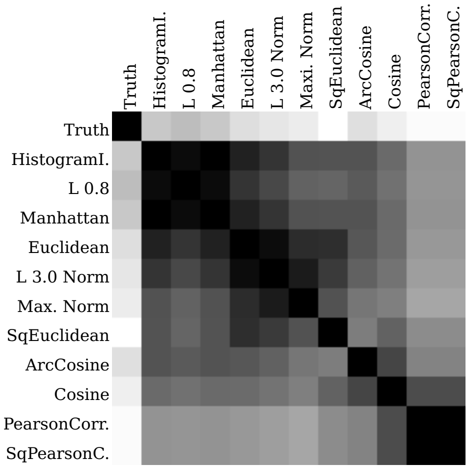


Figure 4: ALOI: LOF, $k = 20$, distance measures

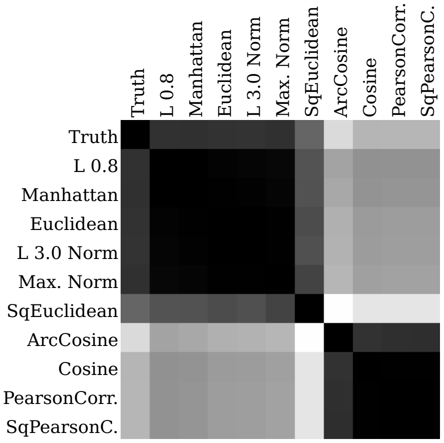
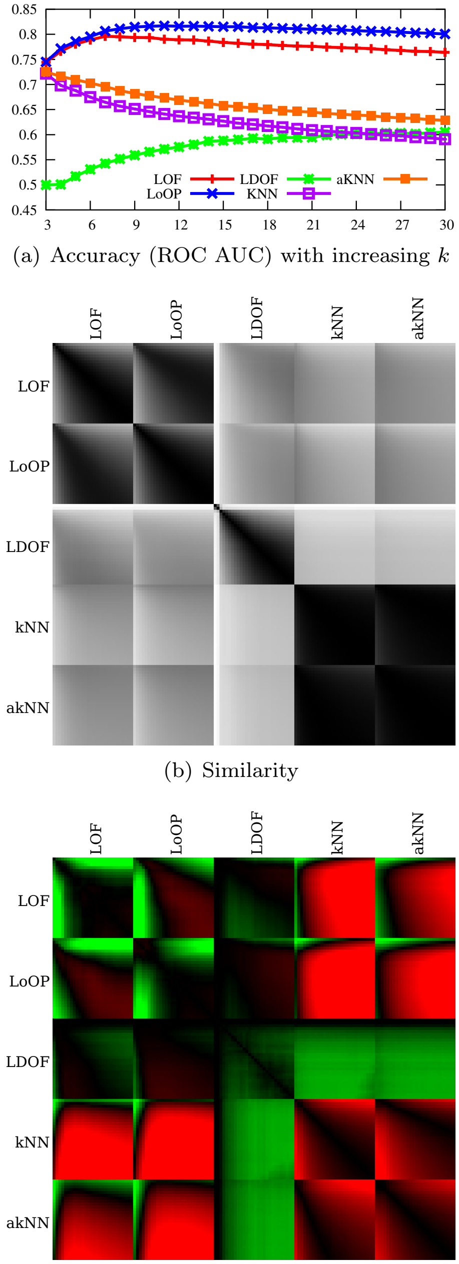


Figure 5: WBC: LOF, $k = 20$, distance measures

5.4 Distance Measures To assess the impact of distance measures on the resulting outlier decisions, we keep the method (again LOF) and parameter ($k = 20$) the same, but vary the distance functions. Used distance measures are histogram-intersection, $L_{.8}$, Manhattan, Euclidean, L_3, maximum, squared Euclidean, arc cosine, cosine, Pearson-correlation, and squared Pearson-correlation. As depicted in Figure 4, on the ALOI data, these distances perform rather differently (though, again, all results are far from perfect). Intriguingly, all results are more similar to each other than they are to the ground truth. This figure changes when we consider the WBC data (Figure 5): here virtually all L_p-norms perform well (and similar), while the vector-length invariant distance measures deteriorate. Overall, we see here again that the WBC data are much more accessible to outlier detection than the ALOI data. Furthermore, we see that the behavior of distance functions



(a) Accuracy (ROC AUC) with increasing k

(b) Similarity

(c) Gain (green: improved, red: deteriorated)

Figure 6: ALOI: Similarity of methods and gain in combining pairs

varies strongly dependent on the data set. The variance in performance of L_p-distances on ALOI is much higher than on WBC data, and there – unsurprisingly – is no clearly "best" distance function. Choosing a different distance function can both have a negligible impact on the algorithms performance or produce largely uncorrelated results.

Table 3: Ensemble results for combination of methods

ROC	gain	combined methods	correl.
0.7218	-	kNN $k = 3$	-
0.7663	-	LOF $k = 4$	-
0.7716	-	LoOP $k = 4$	-
0.7767	-	LOF $k = 20$	-
0.8007	-	LoOP $k = 30$	-
0.8253	0.2176	LOF $k = 20$ + LoOP $k = 4$	0.4006
0.7952	0.1237	LOF $k = 4$ + kNN $k = 3$	0.4226
0.7938	0.0769	LOF $k = 20$ + kNN $k = 3$	0.5014
0.8275	0.1344	LOF $k = 4$ + LoOP $k = 30$	0.5373
0.7814	0.0427	LOF $k = 4$ + LoOP $k = 4$	0.8458
0.7932	-0.0375	LOF $k = 20$ + LoOP $k = 30$	0.9311
reference: existing ensemble methods			
0.7541	mixture model mean [13]		
0.7546	maximum rank [28]		
0.7709	unscaled mean [28]		
0.7821	sigmoid mean [13]		
0.7997	unified score [25]		
0.8054	HeDES scaled mean [35]		

5.5 Diversity vs. Accuracy for Combinations

To numerically measure the gain in outlier detection performance of a combination (ensemble) of outlier detectors as compared to their individual performance, we use the relative improvement towards the target AUC score of 1 over the best of the combined detectors:

$$\text{gain}(M_1, M_2) := 1 - \frac{1 - AUC(M_1 + M_2)}{1 - \max\left(AUC(M_1), AUC(M_2)\right)}$$

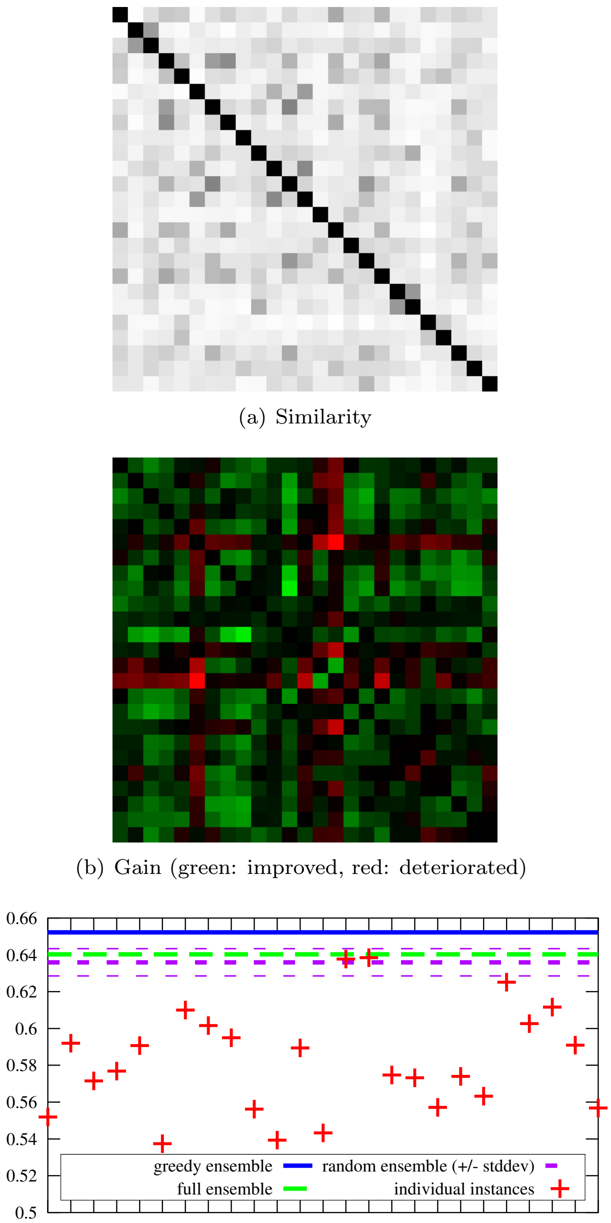
In Figure 6, we inspect the relationship between accuracy of the algorithms (Fig. 6(a)), similarity of their results (Fig. 6(b)), and accuracy gain in combining two algorithms (Fig. 6(c)), here on the ALOI data, varying k for all algorithms from 3 to 30. We see essentially that combining correlated algorithms does not yield a better result than either of them (black) or, depending of the individual performance, even deteriorate (red). The kNN and aggregated kNN models perform well with rather small k. In this region, combination with, e.g., LOF shows good improvements (green). While the accuracy of kNN and aggregated kNN is decreasing with bigger values for k, also the combination with other methods does usually deteriorate the ensemble performance (red). An interesting exception is LDOF, which is substantially different from the kNN methods. Here, the combination of two not extremely well performing methods results in an intriguing gain of performance (albeit at a low level).

Finally, we show numerically that we yield better results when using the correlation information for ensemble member selection as compared to the state of the art as reported in [25] (see Table 3). As a combination rule, we use just the averaged score and rank these. Clearly, this is a rather simple approach. Also, combining just two individuals does usually not improve as well as the combination of more individuals. But we are not interested here in studying combination rules (where the other approaches are more sophisticated already) but in the impact of the ensemble member selection. First, we show five selected base results (among the best performing) along with their achieved ROC AUC. Second, we show for several pairs the ROC AUC and gain achieved when combining two of these, along with their correlation. Clearly, combining highly correlated results achieves a lower gain than combining rather uncorrelated results. Finally, we show as a reference the ROC AUC values for all existing ensemble methods, combining LOF, LDOF, kNN, and agg. kNN as ensemble members. The best reference value is easily reached or surpassed with combinations of two uncorrelated methods, whereas the combination of two strongly correlated methods results in a negative gain. This highlights the importance of diversification for ensembles.

5.6 Effective Ensemble Member Selection We study the impact of correlation analysis on the construction of outlier detection ensembles, inducing diversity with two variants, namely (i) feature bagging, and (ii) using different methods and parameters.

Feature Bagging: In Figure 7, we study the potential of feature bagging for outlier ensembles on the ALOI data. We randomly generated 25 feature bags, each containing 18 out of the 27 original features. In Figure 7(a), we have the similarity matrix of the LOF outlier rankings in these feature-bags. Almost all of them are rather uncorrelated, the highest correlation occurring (except on the diagonal) is 0.477 (bags 7 × 12). As expected, the gains (Fig. 7(b)) are positive (green signal) in most cases. Exceptions are combinations of the best-performing individual feature-bags. These deteriorate (red signal) by pairing with most other feature-bags. We ascertain here that feature bagging usually results in uncorrelated outlier rankings (which qualifies feature bagging a good heuristic for outlier ensembles), and that the combination (at least of more than two) of such uncorrelated rankings usually yields a positive gain.

In Figure 7(c), the individual performance of LOF ($k = 10$) on these feature bags (measured with traditional ROC AUC) is plotted along with the performance of two ensembles resulting from combination of all feature bags. The full ensemble is just the mean score of all instances. It already outperforms the best individual feature bag. The greedy ensemble was built as described in Section 4.3, by collecting the union of the top 250 data points of each instance and using this for weighting and the $0 - 1$ vector as optimization target. The ensemble kept 13 of the 25 feature bags.


(a) Similarity

(b) Gain (green: improved, red: deteriorated)

(c) Accuracy (ROC AUC) of single feature-bags and ensembles

Figure 7: ALOI: Similarity of feature-bags and gain in combinations

For comparison, we computed 5000 truly random ensembles also containing 13 instances each. The mean AUC for this control group is 0.6359, which is slightly below both the best individual method's AUC of 0.6385 and the full ensemble's AUC of 0.6403. The standard deviation is 0.0074, which makes the optimized ensemble's AUC of 0.6522 about 2.20 standard deviations bet-

Table 4: Ensemble results for combination of methods

Method	AUC	significance	gain compared to	
			full	random
Metabolic dataset ($5 \cdot 13 = 65$ instances, $k = 100, 125, \ldots, 400$)				
Full ensemble	0.9201	n/a	:= 0	+56.6%
Random ensemble	0.8159	±0.1221	−130%	:= 0
Greedy ensemble	0.9530	$= \mu + 1.12\sigma$	+41.2%	+74.5%
Pen digits dataset ($6 \cdot 98 = 588$ instances, $k = 3 \ldots 100$)				
Full ensemble	0.9656	n/a	:= 0	+74.6%
Random ensemble	0.8648	±0.1669	−293%	:= 0
Greedy ensemble	0.9697	$= \mu + 0.63\sigma$	+11.8%	+77.6%
ALOI images dataset ($5 \cdot 28 = 140$ instances, $k = 3 \ldots 30$)				
Full ensemble	0.7903	n/a	:= 0	+2.36%
Random ensemble	0.7853	±0.0222	−2.42%	:= 0
Greedy ensemble	0.8380	$= \mu + 2.37\sigma$	+22.7%	+24.6%
KDDCup 1999 dataset ($5 \cdot 10 = 50$ instances, $k = 5 \ldots 50$)				
Full ensemble	0.8861	n/a	:= 0	+15.3%
Random ensemble	0.8655	±0.0414	−18.1%	:= 0
Greedy ensemble	0.9472	$= \mu + 1.97\sigma$	+53.6%	+60.7%

ter than the random subset, and a relative gain of 3.8% over the best individual AUC.

However, in this example, the quality of all individual outlier rankings is not good enough to yield a particularly impressive overall quality of the ensemble (which qualifies the findings of [28]). This is not unexpected on this data set: in color histograms, many dimensions are zero, and few dimensions carry the actual signal. The quality of a feature bag largely depends on keeping the right dimensions.

Combining Different Methods: For our largest experiments, we used all results from LOF, LDOF, LoOP, kNN, and weighted kNN, (for PenDigits also ABOD) and multiple values of k. Some of these instances, such as kNN, are expected to be strongly correlated. Furthermore, LDOF often did not work too well. As control group, we used 5000 random ensembles with as many methods as the greedy ensemble used (usually around $5-10$), and use the mean and standard deviation of these ensembles to judge the significance. It can be seen (Table 4) that the random ensembles are usually clearly worse than the full ensemble, while the greedy ensemble provides a clear improvement, and also performs better than the full ensemble (the PenDigits data set is fairly easy, a much bigger improvement was not to be expected). On the ALOI data set, the ensemble obtained a ROC AUC score of 0.8380, which is in particular also clearly better than the previously reported results on this data set (see Table 3).

6 Conclusion

Though developing algorithms for outlier detection found a lot of attention in recent years, evaluation and judging the quality of algorithms has barely improved over using ROC curves or the area under these curves as a ranking evaluation measure. ROC or ROC AUC analysis, however, completely neglects the scores (which is a pity as long as the scores have any meaning at all) and does not allow for an easy and clear comparison between different methods or for an assessment of similarity between their results.

Here, we developed a correlation measure for comparing rankings, taking the scores into account. This measure is suitable in particular for outlier rankings and outlier scores (posing other requirements as compared to the evaluation of rankings in IR and databases in general). We discussed the relationship and differences to other typical ranking evaluation measures and demonstrated the applicability and advances in evaluating the results of different outlier detection algorithms, with different parametrization, and on different data. Were all methods would have a similar performance as judged by ROC AUC, supplementing a ROC analysis with our ranking similarity measure allows for deeper insights in similarities and differences between different methods, parametrization, distance measures, and data.

In particular, our measure provides for the first time the means to select members of an ensemble for outlier detection in a sophisticated manner (rather than the mere heuristic compositions tried so far). We therefore expect to enable further advances in the development of effective ensemble methods for outlier detection.

References

[1] E. Achtert, A. Hettab, H.-P. Kriegel, E. Schubert, and A. Zimek. Spatial outlier detection: Data, algorithms, visualizations. In *Proc. SSTD*, pages 512–516, 2011.

[2] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *Proc. PKDD*, pages 15–26, 2002.

[3] A. Asuncion and D. J. Newman. UCI Machine Learning Repository, 2007. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[4] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. KDD*, pages 29–38, 2003.

[5] D. C. Blest. Theory & methods: Rank correlation — an alternative measure. *Austral. & New Zealand J. Statist.*, 42(1):101–111, 2000.

[6] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. SIGMOD*, pages 93–104, 2000.

[7] B. Carterette. On rank correlation and the distance between rankings. In *Proc. SIGIR*, pages 436–443, 2009.

[8] T. de Vries, S. Chawla, and M. E. Houle. Finding local

anomalies in very high dimensional space. In *Proc. ICDM*, pages 128–137, 2010.

[9] T. G. Dietterich. Ensemble methods in machine learning. In *Proc. MCS*, pages 1–15, 2000.

[10] R. Fagin. Combining fuzzy information from multiple systems. *JCSS*, 58:83–99, 1999.

[11] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

[12] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *JCSS*, 66(4):614–656, 2003.

[13] J. Gao and P.-N. Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Proc. ICDM*, pages 212–221, 2006.

[14] J. M. Geusebroek, G. J. Burghouts, and A. Smeulders. The Amsterdam Library of Object Images. *Int. J. Computer Vision*, 61(1):103–112, 2005.

[15] J. Ghosh and A. Acharya. Cluster ensembles. *WIREs DMKD*, 1(4):305–315, 2011.

[16] U. Güntzer, W.-T. Balke, and W. Kießling. Optimizing multi-feature queries for image databases. In *Proc. VLDB*, pages 419–428, 2000.

[17] A. S. Hadi, A. H. M. Rahmatullah Imon, and M. Werner. Detection of outliers. *WIREs Comp. Stat.*, 1(1):57–70, 2009.

[18] A. M. Ivanescu, M. Wichterich, and T. Seidl. ClasSi: measuring ranking quality in the presence of object classes with similarity information. In *Proc. PAKDD Workshop QIMIE*, 2011.

[19] W. Jin, A. Tung, and J. Han. Mining top-n local outliers in large databases. In *Proc. KDD*, pages 293–298, 2001.

[20] F. Keller, E. Müller, and K. Böhm. HiCS: high contrast subspaces for density-based outlier ranking. In *Proc. ICDE*, 2012.

[21] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

[22] W. Klement, P. Flach, N. Japkowicz, and S. Matwin. Smooth receiver operating characteristics (smROC) curves. In *Proc. ECML PKDD*, pages 193–208, 2011.

[23] E. M. Knorr, R. T. Ng, and V. Tucanov. Distance-based outliers: Algorithms and applications. *VLDB J.*, 8(3–4):237–253, 2000.

[24] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proc. CIKM*, pages 1649–1652, 2009.

[25] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proc. SDM*, pages 13–24, 2011.

[26] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proc. KDD*, pages 444–452, 2008.

[27] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proc. WWW*, pages 571–580, 2010.

[28] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proc. KDD*, pages 157–166, 2005.

[29] B. Liebl, U. Nennstiel-Ratzel, R. v. Kries, R. Fingerhut, B. Olgemöller, A. Zapf, and A. A. Roscher. Very high compliance in an expanded MS-MS-based newborn screening program despite written parental consent. *Prev. Med.*, 34(2):127–131, 2002.

[30] T. A. Maturi and E. H. Abdelfattah. A new weighted rank correlation. *J. Math. & Stat.*, 4(4):226–230, 2008.

[31] M. Melucci. On rank correlation in information retrieval evaluation. *ACM SIGIR Forum*, 41(1):18–33, 2007.

[32] E. Müller, M. Schiffer, and T. Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *Proc. ICDE*, pages 434–445, 2011.

[33] S. Nepal and M. V. Ramakrishna. Query processing issues in image (multimedia) databases. In *Proc. ICDE*, pages 22–29, 1999.

[34] R. Newson. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*, 2(1):45–64, 2002.

[35] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *Proc. DASFAA*, pages 368–383, 2010.

[36] G. H. Orair, C. Teixeira, Y. Wang, W. Meira Jr., and S. Parthasarathy. Distance-based outlier detection: Consolidation and renewed bearing. *PVLDB*, 3(2):1469–1480, 2010.

[37] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proc. SIGMOD*, pages 427–438, 2000.

[38] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *Amer. Statistician*, 42(1):59–66, 1988.

[39] G. S. Shieh. A weighted Kendall's tau statistic. *Stat. & Prob. Letters*, 39(1):17–24, 1998.

[40] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proc. SIGIR*, pages 66–73, 2001.

[41] C. Spearman. The proof and measurement of association between two things. *Amer. J. Psychol.*, 15(1):72–101, 1904.

[42] M. Swain and D. Ballard. Color indexing. *Int. J. Computer Vision*, 7(1):11–32, 1991.

[43] G. Valentini and F. Masulli. Ensembles of learning machines. In *Proc. Neural Nets WIRN*, pages 3–22, 2002.

[44] N. H. Vu and V. Gopalkrishnan. Efficient pruning schemes for distance-based outlier detection. In *Proc. ECML PKDD*, pages 160–175, 2009.

[45] S. Wu and F. Crestani. Methods for ranking information retrieval systems without relevance judgements. In *Proc. ACM SAC*, pages 811–816, 2003.

[46] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proc. SIGIR*, pages 587–594, 2008.

[47] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proc. PAKDD*, pages 813–822, 2009.