Ensembles for Unsupervised Outlier Detection: Challenges and Research Questions

[Position Paper]

Arthur Zimek
Ludwig-Maximilians-Universität
Munich, Germany
http://www.dbs.ifi.lmu.de
zimek@dbs.ifi.lmu.de

Ricardo J. G. B. Campello
University of São Paulo
São Carlos, Brazil
http://www.icmc.usp.br
campello@icmc.usp.br

Jörg Sander
University of Alberta
Edmonton, AB, Canada
https://www.cs.ualberta.ca
jsander@ualberta.ca

ABSTRACT

Ensembles for unsupervised outlier detection is an emerging topic that has been neglected for a surprisingly long time (although there are reasons why this is more difficult than supervised ensembles or even clustering ensembles). Aggarwal recently discussed algorithmic patterns of outlier detection ensembles, identified traces of the idea in the literature, and remarked on potential as well as unlikely avenues for future transfer of concepts from supervised ensembles. Complementary to his points, here we focus on the core ingredients for building an outlier ensemble, discuss the first steps taken in the literature, and identify challenges for future research.

1. INTRODUCTION

Outlier detection is the process of identifying those observations which deviate substantially from the remaining data. Many definitions of outliers exist in the statistics literature, usually tied to specific assumptions on the underlying data distribution. The most common general definitions remain rather vague, such as these classic examples:

> "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism." [30]

> "An outlying observation, or 'outlier,' is one that appears to deviate markedly from other members of the sample in which it occurs." [26]

> "An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [8]

The point of all these definitions is the idea that any process, whether it is a traffic network, web server traffic, credit card data, sensor data in some scientific experiment, or the human metabolism, offers characteristic observations that could even be predicted if the process was well-understood. Any unpredicted observation indicates a lack of understanding of the particular process, or is produced by a different process (such as a traffic accident, a network intrusion attack, credit card fraud, sensor failure, or a disease affecting human health), and therefore probably is worth further investigation.

Outlier detection algorithms aim to automatically identify those valuable or disturbing observations in large collections of data. Because there is no rigid definition of which observation exactly is an outlier, every algorithm is based on a model that is relying on certain assumptions of what qualifies as an outlier. Clearly, the applicability of each model depends on the nature of the data. Sophisticated algorithms do not only label observations as outlier or inlier, but assign scores to observations, representing degrees or probabilities of outlierness. Some popular models are based on the distance between objects [37; 60; 4; 74], or on the density of the neighborhood of an object [9; 56; 34; 38; 42], or based on the variance of angles between object vectors [42; 58], or on other principles of outlierness in various domains [12; 13; 3]. These methods represent different attempts to make the rather vague intuition about what outliers are more concrete, typically in an implicit, procedural way [65].

Because every model is specialized for different characteristics of observations and therefore fits only to some aspects of the "whole truth", it might be a good idea to integrate various different outlier detection results, producing a consensus of judgements. The key idea of such an approach, which is called an "ensemble", is that the combination of individual judgements, or outlier detection results, is beneficial if those judgements do not contain all the same errors. One might think of it as a majority vote of a jury (as in Condorcet's Jury theorem [47]): One or another judgement about an observation might be wrong, but the majority might still be right, as long as the judgements are, overall, somewhat reliable and every member decides independently from the others.

Aggarwal [2] recently proposed a categorization of ensemble approaches to outlier detection by algorithmic patterns or strategies. He distinguishes "sequential ensembles" vs. "independent ensembles", and "model-centered ensembles" vs. "data-centered ensembles". This is helpful for identifying aspects of the ensemble approach in the literature. Accordingly, he points out that before the first paper was explicitly talking about "outlier ensembles" [45], traces of the very idea of combining different models have appeared earlier in the literature, and also several times later without discussing a potential relationship to ensemble techniques explicitly. When reading the literature through these glasses of ensembles, we can undoubtedly find many hints on the ensemble idea without explicit discussion. However, not everybody has to wear these glasses. To discuss the problem

of, e.g., subspace outlier detection based on the combination of several models [36; 51; 52] without discussing ensemble techniques is perfectly fine. In fact, the subspace outlier problem is a hard problem in its own right and the typical conference paper cannot accommodate a broader discussion for reasons of space restrictions.

Furthermore, the subspace outlier problem could be seen as a problem analogous to the multiview or alternative clustering problem [77] where it is not intended to find the consensus clustering; instead, different clustering solutions in different subspaces can each be interesting, valid solutions. Likewise, different outliers in different subspaces could each be meaningfully reported. This is reflected in recent research addressing the explanation of subspace outliers [14]. Seen this way, subspace outlier detection would even be orthogonal to the "ensemble" or "consensus" idea.

Nevertheless, discussing the subspace outlier problem while taking into account reasoning on ensemble techniques would seem promising of finding more principled solutions to the subspace outlier problem [76]. Likewise, it would seem that ensemble techniques such as feature bagging [45], i.e., using different subspaces as a means to learn diverse models, could also benefit from insights in the area of subspace outlier detection.

Complementary to Aggarwal [2], we would like to discuss here the specific challenges, the first steps taken so far in the literature, and overall the important questions in research regarding ensembles for outlier detection.

Transferring basic principles from supervised learning, the two key principles of ensemble construction would be *accuracy* and *diversity*. Casting outlier detection as an unsupervised problem, however, there is nothing known about the *accuracy* of individual outlier detectors during learning. This is a very fundamental problem and, as Aggarwal [2] pointed out, probably one of the main reasons why the state of the art in research on ensembles for unsupervised outlier detection is not very advanced. But obviously this problem would also affect ensemble clustering where we have a lot more of research presented in the literature. Therefore, we should have a closer look on the differences between ensemble clustering and ensemble outlier detection beyond their common characteristic of being unsupervised ensembles. How to assess the *diversity* of outlier detection results does not have a straightforward answer either, but at least it found some attention recently.

In the remainder of this paper, we will first have a look at the research area of ensemble clustering in Section 2, detailing why ensembles for outlier detection are quite a different issue. We will discuss the crucial research questions for outlier detection ensembles, reflecting the literature as sparse as it is so far, in Section 3, Section 4, and Section 5. Common approaches to assess the accuracy of outlier detection results are far from satisfying. We sketch the problem in Section 3. The diversity of models, besides their accuracy, is the most important ingredient for ensemble construction. We will discuss the issue of diversity of models for outlier detection in Section 4. Another central question is how to actually construct the ensemble, i.e., how to combine the models. The challenges in combining different models and preliminary findings in the literature will be discussed in Section 5. Finally, we summarize our positions in Section 6.

2. ENSEMBLE CLUSTERING, ENSEMBLE OUTLIER DETECTION — WHAT IS THE DIFFERENCE?

Using ensemble techniques to improve classification is based on a sound theory [16; 70; 10; 44; 62]. In the unsupervised area of clustering, using ensemble techniques has at least a history of many empirical studies [67; 25; 55; 24; 33]. Furthermore, the idea of using several different clustering results is important not only in ensemble clustering as an explicit technique but also in related approaches such as multi-view clustering, subspace clustering, and alternative clustering [11; 59; 31; 50; 77]. The ensemble idea has also been used when clustering evaluation measures are combined [72].

By a simple transfer of ideas from these research results in the area of ensemble clustering (and related areas), we can assume that a combination of outlier detection models would also show potential to improve considerably over the combined individual models. Also, we can assume, by analogy, that diversity of models would be helpful in outlier detection as it is in clustering or classification.

Surprisingly, for outlier detection there have not been many attempts to use ensemble techniques for improvement in a principled way, let alone investigations of the theoretical basis of doing so. When comparing the tasks of outlier detection and clustering, we can name several reasons for this surprising fact — reasons, that, at the same time, highlight the research issues that are different for the design of ensemble methods for outlier detection than for ensemble clustering.

1. The first issue is the question of how to measure *accuracy* (or some other index of *quality*) of the result of an unsupervised data mining task. In the case of clustering, we distinguish external and internal validity measures.

   - External measures would assess how well some clustering reflects an externally given ground truth (gold standard) partition, using similarity measures such as the Rand-index [61] or the adjusted Rand-index (ARI) [32], or other pair counting approaches [57], essentially counting the number of agreements and disagreements regarding the membership of pairs of objects to clusters in the ground truth and the clustering solution. Other similarity measures compare partitions by mapping sets, important examples being entropy-based measures such as (normalized) mutual information [67; 48].

     Although the approach to assess clustering quality by comparison with some given ground truth of known classes is debatable [22], there is probably no better approach available to assess clustering quality w.r.t. external knowledge.

   - Internal measures evaluate the quality of some clustering result according to certain assumptions on what constitutes a good clustering, for example, compactness and separation of clusters as, e.g., in the Silhouette coefficient [35], or density-connectivity within and density-separation between clusters as, e.g., in DBCV [49]. Many of these internal measures can also be used to rank different solutions relatively to each other and are therefore

also called "relative validity measures" [71]. This way, potential members for a clustering ensemble could be selected based on their relative quality, regardless of any information on the ground truth [20; 53].

Quality assessment for outlier detection models is quite different since outlier detection models are not partitioning but ranking the data. We will elaborate on variants of quality assessment in outlier detection in Section 3.

2. The second issue that is important for building good ensembles but, at the same time, is quite different for outlier detection and for clustering, is the *diversity* of models. For clustering, assessment of diversity again can make use of numerous similarity measures such as external validity measures. The effect of diversity of models, and different ways of designing better ensembles making use of diverse components, has been studied extensively in the literature on ensemble clustering [43; 28; 27; 20; 6].

The similarity of outlier detection models, again, not being partitions but rankings of the data, cannot rely on the same means. We will discuss first approaches addressing this aspect of ensembles for outlier detection in Section 4.

3. The third issue, when given individual models (that are, hopefully, accurate and diverse), is how to *combine* these models. For ensemble clustering, this requires some matching of partitions, e.g., using similarity measures (again, as those available from external validity measures) or more refined methods of deriving some consensus partition [67; 19; 5; 69; 68; 21; 25; 55].

The combination of outlier detection results, i.e., rankings, requires different techniques than the combination of partitions studied in ensemble clustering. We discuss the issues particularly involved in combining outlier detection models and the approaches presented so far in the literature in Section 5.

3. ACCURACY (QUALITY) OF RESULTS

3.1 External Evaluation

If given a ground truth dataset where we know, for each object, whether it actually is an outlier or not, two ways of measuring the quality of the outlier detection result are commonly used in the literature [76].

The first, more widely used measure of success is based on receiver operating characteristic (ROC) curves. ROC curves plot the true positive rate against the false positive rate. The resulting, monotone curves are usually turned into a measure by computing the area under this curve (AUC). This allows to display several results in a single graph and to compare the results numerically.

For a random ranking result, both rates (true positive rate and false positive rate) will grow at the same rate, resulting in an area that approximately fills half of the space. For a perfect result, returning all outliers first and only then returning the inliers (i.e., we have 100% true positives before we even get the first false positive), the area under the corresponding curve will cover the available space completely,

i.e., the maximal ROC AUC value is 1.0. Intuitively, the ROC AUC value can be seen as the probability that a pair of two randomly chosen objects, one positive example (outlier) and one negative example (inlier), is sorted correctly (i.e., the outlier is ranked before the inlier) [29]. ROC curves and ROC AUC analysis inherently treat the class imbalance problem by using the relative frequencies which makes them particularly popular for evaluation of outlier detection.

Sometimes, additionally or alternatively to ROC analysis, the precision of the result is assessed for a given number k of top outliers: How many of the top k ranked data objects are actually outliers? This is known as "precision at k". As an evaluation measure, this is a bit more problematic, as it involves a parameter.

Both quality measures require data with known, annotated outliers, or, to put it in terms of classification, a binary, yet typically highly imbalanced classification task (very few outliers vs. many inliers).

Although the task of outlier detection is practically ubiquitous, these practical tasks are tasks because the ground truth is unknown. There is nothing like established benchmark data sets in this field, required to study and compare the behavior of algorithms for outlier detection. What people do, for example, is using classification data sets such as available in the UCI repository [7]. To prepare an outlier detection task from such classification tasks, one can, e.g., pick some class as outlying and keep only a small sample of this outlier class while the other classes remain complete and are treated as inliers. This procedure is sometimes called "down sampling" and has been used, with different variants, in many studies designing new outlier detection methods [1; 73; 42; 74; 36; 15; 14]. A recent study is dedicated to develop a more systematic approach [18], but this is also merely a wider step in the same direction — a direction that probably is debatable.

Let as note, however, that all these problems regarding external evaluation are not specific for outlier detection *ensembles* but are inflicting the research on outlier detection in general.

3.2 Internal Evaluation

To the best of our knowledge, there are no insights whatsoever in the literature on outlier detection regarding internal validation measures.

3.3 Challenges

As challenges for future research on the aspect of quality assessment of outlier detection results, we see the following issues and questions for research:

- Defining useful, publicly available benchmark data in a principled way for outlier detection tasks would allow for a more objective study of progress in the field. This is, however, a general challenge in outlier detection not restricted to improving the field on ensemble methods for outlier detection.

- It will be very useful to identify meaningful criteria of internal evaluation of outlier detection rankings. Again, this is important for the progress of research on outlier detection in general. But internal evaluation criteria can be expected to have significant impact in particular on the research in ensemble methods for

outlier detection, as this might allow to develop similar ideas for outlier detection ensembles as mentioned above for better clustering ensembles based on relative validity criteria.

Actually, since the effectiveness of internal criteria themselves will need to be evaluated, the first challenge posed above, namely, to provide better, more principled and objective possibilities for external evaluation of results, will be an important prerequisite for the second challenge.

4. DIVERSITY OF MODELS

Typical unsupervised methods for outlier detection return a score of "outlierness" for each object. When assuming a fixed order of the objects in the dataset, an outlier detection result can be thought of as a vector consisting of the outlier scores for each object [63]. This way, we can define the space of all possible outlier detection results for a given dataset, where each dimension represents the possible outlier scores of a particular observation. Let us, for the sake of simplicity, consider some outcomes for two objects, resulting in a two-dimensional plot as illustrated in Figure 1 (for n objects, the space would be n-dimensional and not suitable for visualization). The green circle represents the (usually unknown) ground truth, while the red crosses are individual results generated by somehow diverse outlier models.

Figure 1(a) shows how the combination of the six individual result vectors by the simple component-wise mean produces another result, represented by the blue X. This combined result is, in this case, a better approximation of the ground truth (green circle). All individual solutions are already quite accurate, that is, they are close to the ground truth. This is a necessary condition for assembling these individual solutions to make a good ensemble, which can be illustrated by the following reasoning: It is known that the ground truth is located somewhere in the result space, but it could be anywhere. The generation of multiple individual, more or less accurate (i.e., at least better than random) results restricts the space of where the true result most probably lies: if they are accurate to some extent, the true result will be close to them. The motivation for combining the individual results by, for example, computing the mean score for each observation is the expectation that the true result will be somewhere between them. In fact, for combination techniques like the mean, the convex hull of the individual results already restricts the result space to an area where the true result is expected to be, and where ensembles generate their integrating results.

Figure 1(b) illustrates the limited effects of accuracy when diversity is missing. It can easily be seen that, again, the individual results are quite accurate. However, the combined result gets rather attracted towards the majority of the single results. If the upper right, rather deviating, result would not exist, the combined result would lie completely inside the tight cluster of remaining results and would be even more distant to the true result. This is the effect of missing diversity. All single results of that tight cluster make the same error: They underestimate the outlier score for the object that is represented by the x-axis. In comparison, Figure 1(a) shows results which make different errors, each of them over- and underestimating a score, resulting in an accurate ensemble result.
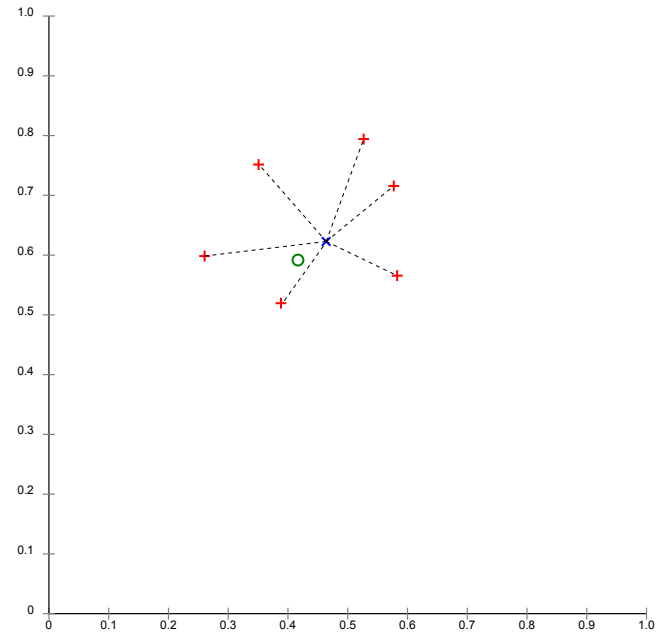
(a) Diverse results
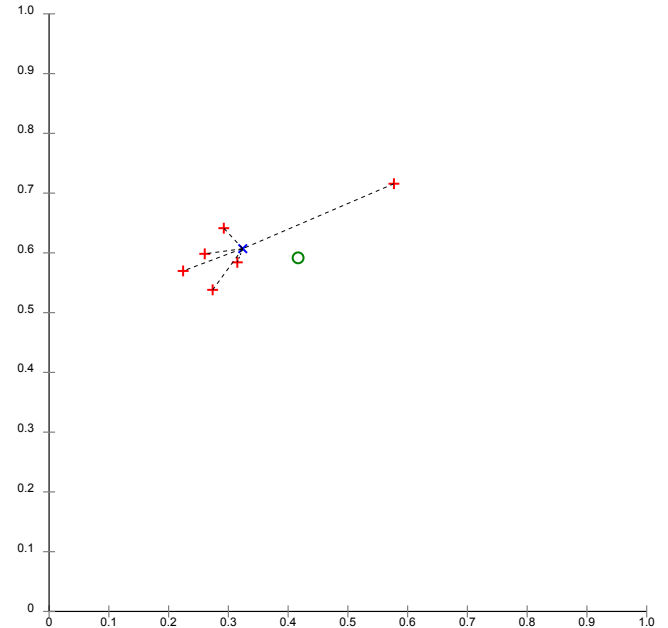
(b) Clustered results

Figure 1: Diverse and clustered outlier detection results.

Of course, diversity is not the only criterion. Ignoring accuracy and maximizing diversity would scatter the individual results all across the result space without any restriction. Both the true result and a combined result could reside anywhere in the complete result space, not necessarily being close to each other.

Seeing the outlier detection results as vectors in a vector space, spanned by the observations of a given dataset, as depicted in Figure 1, allows us to see both components, accuracy and diversity: accuracy of individual ensemble members (red crosses) is represented by the *absolute distances*

from the true result (green circle), while the diversity is reflected in the *relative distances*, taking into account also *the direction of deviation*, from the true result. Clearly, both criteria, accuracy and diversity, are antagonistic to a certain degree. The more accurate the individual results are, the tighter they are packed and therefore the less diverse they are. And the more diverse the individual results are, the less accurate most of them can possibly be. If we just transfer the intuition on ensemble learning from supervised ensembles, the essential requirement is that individual ensemble members would commit errors different from the other ensemble members whenever they are committing errors at all. As long as they are correct they also should be in accordance with the other ensemble members. Let us note that, interestingly, the vector space intuition of outlier detection results would also allow us to talk about subspaces (i.e., subsets of objects) where result vectors would cluster. So maybe it is possible to transfer insights from subspace clustering [41; 66; 77] to the area of outlier ensembles.

So far, this is an intuition basically transferred from supervised ensemble learning without any further theoretical understanding. Questions tackled in the literature are how to induce diversity among models (Section 4.1) and how to assess diversity along with the impact of diversity on the ensemble performance (Section 4.2). If we are given diverse models, another interesting question is if we can select some of them to build better ensembles than if we selected all of them. We sketch a greedy approach to this problem (Section 4.3) and suggest challenges and issues for future research w.r.t. to the diversity of models (Section 4.4).

4.1 Inducing Diversity

Mostly in analogy to methods for inducing diversity in classification ensembles [10] or clustering ensembles [24], there have been studies on combining outlier scores (or rankings) (1) learned in different subsets of attributes (i.e., different subspaces), (2) learned on different subsets of objects, (3) learned by randomized methods, (4) learned by the same method but using different parametrization, and (5) learned by different models.

1. Combining outlier scores or rankings learned on different subsets of attributes, the so-called "feature bagging" was the first paper to explicitly discuss building ensembles for outlier detection [45].

   Some meta methods that specialize on subspace outlier detection, such as HiCS [36] or OutRank [51], do in fact also combine models learned in different subspaces. However, their subspaces are not selected randomly but from the point of view of detecting *subspace* outliers. From an ensemble point of view this restriction could be a disadvantage by introducing a bias. This bias is meaningful and of course intended under the assumption of noisy subspaces that ought to be excluded from the feature bagging procedure. Although these subspace methods are not discussed as an ensemble framework, in effect they could be understood and probably discussed more deeply in an ensemble context, as pointed out earlier [76; 2].

2. The orthogonal approach, combining models learned on different subsets of objects [75], has the advantage of a considerable speed-up as it works well in particular with small sample rates, e.g., using only 10% of the objects for the density estimates. The typical complexity of outlier detection methods is in $\mathcal{O}(n^2)$ due to kNN queries. Hence a common ensemble, such as feature bagging, would be in $\mathcal{O}(s \cdot n^2)$ for s ensemble members. The subsampling ensemble, however, has to perform, for each data object (n), a kNN query on a subsample only (i.e., $m \cdot n$ for sample rate $0 < m < 1$). Repeated on s subsamples, this results in $\mathcal{O}(n \cdot mn \cdot s)$. For example, with a sample rate of 10% and an ensemble size of 10 members, the ensemble requires roughly the same runtime as a single base learner on the full data set while the runtime of a standard ensemble (combining models learned on the complete dataset) would be roughly 10 times the base learner's runtime.

3. The approach of "isolation forests" [46] designs a randomized method leading to diversity. The effect of randomized methods is, however, complemented in this method by the effect of random subsamples of the dataset. On each subsample, a binary tree (called "isolation tree") is built splitting randomly selected attributes at randomly selected split points up to a specified depth or until a given subset at some node cannot be split (because it has only one element). The path length for an object in the tree is expected to be shorter in areas of lower density. Hence, the ensemble (the forest) is essentially an aggregate of randomized density estimates.

   The speed-up effect of the subsampling is not as prominent in the approach of isolation forests as in the more general approach for methods based on kNN queries [75] since no kNN queries are required for building the binary trees anyway. A possible effect of using subsamples here is that the trees, because these subsamples are unlikely to contain many outliers, are built faster and describe the dataset in a more concise way. The effect of inducing diversity by drawing subsamples appears to be less important for the success of this method than the effect of the randomization of attribute and split point selection [46, Fig. 18].

4. Combining models learned using different parameters has been proposed as an ensemble method by Gao and Tan [23], although Aggarwal [2] identified traces of this idea in earlier papers without an explicit discussion of the ensemble idea. For example, for LOF [9] as well as for LOCI [56], the authors suggested to try different parameters (controlling the granularity of neighborhood analysis; in LOCI, the combination of different granularity is even inherent to the method) and to use the granularity that optimally enhances the outlierness. Following this model, a recent generalized KDE-based method [64] also combines models learned for different parameters to an ensemble approach.

5. Combining outlier scores of different algorithms (i.e., combinations of different models of what constitutes an outlier) has been explored in several studies [54; 40; 63].

   When combining scores from different models, normalization and unification of the scores that can scale very differently and sometimes are even inverted (i.e., some

methods represent outliers by large scores, some methods represent outliers by small scores) becomes essential (although normalization should be considered for combination of scores from different subspaces or different parameters as well, depending on the properties of the model).

The first of these studies [54] used a generic normalization, the second [40] studied the properties of distributions of scores for different methods and, thus, was able to use specialized normalization procedures. The third [63] proposed a greedy combination procedure based on an assessment of diversity of models.

Overall, these studies highlight three important aspects for outlier ensembles: assessment of diversity, normalization of scores, and combination procedures, which we discuss in Sections 4.2, 5.1, and 5.2, respectively.

The greedy combination strategy [63] also raises an interesting challenge for outlier ensembles: how to choose good ensemble members or how to train improved ensemble members based on previously learned and evaluated models in the absence of a ground truth for evaluation. This challenge has also been noted by Aggarwal [2]. The heuristic of the greedy ensemble will be discussed in Section 4.3.

At first sight, thinking about the transfer of techniques from classification or clustering ensembles, it may seem that with the five mentioned categories of heuristics for inducing diversity the obvious opportunities have been studied in the literature. However, all these studies leave room for deeper understanding of these heuristics and there are probably more methods for inducing diversity waiting to be explored that perhaps do not have a counterpart in classification or clustering ensembles. For example, diverse models could be learned by using different distance measures. This has only been studied partly [63], assessing the resulting diversity of models but not the quality of ensembles combining these models.

4.2 Assessing Diversity

Having seen different methods for inducing diversity that are known from classification or clustering ensembles, the question arises how well these methods work in the context of outlier detection. This question was addressed by a recent study [63], proposing the vector space of outlier scores that we sketched above and weighted Pearson correlation as a similarity measure for these score vectors. This study discussed two use cases of such an assessment of diversity: (1) studying the suitability of methods for inducing diversity and (2) selecting the most diverse models for combination. The latter we discuss in Section 4.3, as mentioned earlier. Let us discuss the first aspect now.

The idea of using a weighted similarity measure, such as weighted Pearson, to compare score vectors, is motivated by the relative importance of outlier scores while differences in inlier scores should not matter that much. Given a ground truth (i.e., using some dataset with known outliers), the weights for the similarity measure comparing score vectors can be adjusted to this ground truth. Studying some methods, some distance measures, and some datasets using such a weighted similarity measure, the findings reported by Schubert et al. [63] are:

- Outlier score vectors are usually similar (strongly correlated) between the same model just using different

parameters (e.g., different values of k for the neighborhood size, i.e., different granularity). Hence, this method of inducing diversity is probably not a good idea for building ensembles as long as the model used is not known to be an "unstable learner" w.r.t. its parameter. The classic models LOF [9], kNN outlier [60], and kNN weight [4], among others, are shown experimentally to be rather stable w.r.t. neighborhood size.

- Different distance measures can have a stronger impact on diversity on some datasets. But there are also examples where all L_p norms result in strongly correlated results while results based on vector-length invariant distance measures (such as the cosine distance and other correlation measures) again are correlated strongly with each other but not with the L_p-norm-based results. Using different distance measures hence seems to be promising although the suitability of some distance measure is highly dataset dependent.

- Different models (algorithms) fall into families that learn similar results. For example, the results of LOF [9] and of the LOF variant LoOP [38] seem highly correlated, and the results of the kNN model [60] and the kNN weight model [4] are strongly correlated as well (on the datasets studied), but LOF or LoOP (both being *local* methods) and kNN or kNN weight (both being *global* methods [65]) are not strongly correlated. Combining models from different families might have a large potential to lead to improved ensembles.

- Feature bagging [45] apparently has the potential to lead to very uncorrelated results and, thus, to improved ensembles.

The finding of weakly correlated results by feature bagging is also reflected in the finding of feature bagging being rather unstable [75]. From the perspective of building ensembles, instability is not necessarily a bad thing although the combination of very different models is not bound to lead to a good ensemble. Diversity, after all, is only one aspect besides accuracy and too much of diversity is bound to limit accuracy.

4.3 Model Selection

Aggarwal [2] pointed out that analogues of "boosting" or "bucket of models" — established concepts for supervised ensemble learning — are unlikely to be developed for unsupervised outlier detection. We pointed out (Section 3), that internal validity measures for outlier detection results are still missing in the literature and would be very important. Yet this does not mean that model selection is impossible: at least a greedy heuristic, optimizing diversity in an ensemble, has been discussed recently [63]. This rather crude heuristic (see a sketch in Algorithm 1) relies on an accuracy *estimate* based on all learned potential ensemble members. The method first takes the union of the top k points of all results as preliminary outliers for determining weights for the similarity measure (weighted Pearson), assessing the diversity between results. Then the ensemble is composed, starting with the result that is closest to this consensus result. Next the remaining outlier detectors are sorted by the *lowest* correlation to the result of the current ensemble (initially, the ensemble consists only of one outlier detector) and test if including the next detector would improve

Algorithm 1: Greedy Model Selection

```
/* individual outlier detectors:         */
I := list of individual outlier detectors;
K := union of top-k outliers ∀I;
/* K are the preliminary 'outliers'       */
v := target vector;
/* (vi = 1 if object i ∈ K, vi = 0, otherwise)   */
E := ∅ ensemble;
sort I by weighted Pearson correlation to v;
/* weights: 1/(2|K|) (outliers), 1/(2(n-|K|)) (inliers) */
E := E ⋃ getFirst(I);
p := current prediction of E;
sort I by weighted Pearson to p (decreasing order);
while I ≠ ∅ do
    i = getFirst(I);
    if weightedPearson(E ⋃ i, v) > weightedPearson(E, v)
    then
        E := E ⋃ i;
        p := current prediction of E;
        sort I by weighted Pearson to p (decreasing order);
    end
end
return E;
```

the correlation of the ensemble result with the (preliminary) target vector (i.e., the *estimated* ground truth). If yes, this detector is included in the ensemble and the list of remaining detectors is reordered. If no, the detector is discarded and the algorithm continues with the next detector. Note that the whole process works in a completely unsupervised manner in that no *actual* ground truth is used.

This heuristic is based on the assumption that the union of the complete set of individual outlier detectors is *somehow* accurate but can be improved by dropping those detectors that are strongly correlated with others. This assumption serves to overcome the limitations of unsupervised ensemble learning by unavailability of training data. Although we cannot see — so far — analogues of boosting for unsupervised learning either, very likely better heuristics than this greedy model selection are possible. In terms of the issue of accuracy (Section 3), this heuristic is using internally constructed means of validation *as if* it were externally available ground truth.

4.4 Challenges

As challenges for future research on the aspect of diversity for outlier ensembles, we see the following research questions:

- A thorough study of the heuristics to induce diversity proposed so far, in more detail and in comparison with each other, would be very interesting. As the heuristics are rather different from each other, a decent and fair comparison study is far from trivial. One could study the diversity actually achieved by the methods for inducing diversity, and the impact of diversity on the performance of an ensemble.

- To identify yet other methods for inducing diversity would probably lead to broader applicability and deeper understanding of the impact of diversity and of the stability of base methods.

- To develop better measures of diversity of outlier score vectors in the absence of ground truth would also be crucial to help us understand the issues of diversity and stability.

- Effective methods of choosing appropriate ensemble members (i.e., those that are different from other ensemble members) would also be desirable. Answers to the previous issues would allow progress here as well.

5. COMBINATION OF MODELS

Having derived a couple of outlier detection results, or vectors of outlier scores, that are – ideally – diverse and accurate to some extent, the third central question is how to combine them to derive a consensus or ensemble result. The two issues we discuss here in particular are the requirement of score normalization for a meaningful combination of scores (Section 5.1) and the different possibilities to combine (normalized) score vectors (Section 5.2). Some prefer to combine the rankings instead of the score vectors which we will touch upon (Section 5.3). We suggest challenges and issues for future research w.r.t. the combination of models (Section 5.4).

5.1 Normalization of Scores

Any meaningful combination of score vectors relies heavily on the scores provided by the individual outlier detectors being comparable. This problem practically rules out the combination of different base methods or, for many methods, different parametrizations of the same method (e.g., different k for a kNN-distance-based method, as the result with the largest k would dominate the distance values). Even when using the same method as base outlier detector and identical parametrization, outlier scores obtained from different subspaces could vary considerably, if some subspaces have largely different scales. The ensemble could then be dominated by just one of the feature bags.

Several of the papers discussing outlier detection ensembles focused on the issue of comparability of scores for score combinations. The first approach was to use sigmoid functions and mixture modeling to fit outlier scores, provided by different detectors, into comparable probability values [23]. The second approach was scaling by standard deviation [54]. Finally, statistical reasoning about typical score distributions by different methods enabled normalizations tailored to particular properties of different methods [40]. Although the solutions provided so far probably leave room for improvements, the important thing is to realize the problem and to use *some* normalization when combining outlier scores. To provide a good estimate of the actual probability of some object being an outlier is something valuable for supporting the user in the interpretation of the individual outlier detection result. For combination of several results into an ensemble, this calibration is perhaps not equally important. But normalization of scores is important, to avoid a bias of the decision to the individual result with the largest scale. This distinction is somewhat analogous to the distinction between *class probability estimates* and *classification decisions* based on these probability estimates that has been emphasized for the understanding of the performance of the naïve Bayes classifier [17]: the outlier scores need not be good absolute outlier "probabilities" in order to make

sense for a combination but their relative scale needs to reflect the actual ratio of outlierness for compared objects.

5.2 Combination of Score Vectors

Let us assume we are provided with accurate, diverse, and normalized outlier scores, where normalization includes regularization [40], i.e., without loss of generality we can assume that the larger score denotes "more" outlierness while inliers should have been assigned small score values. Now the question remains how to combine the scores. Aggarwal [2, Section 4.2] discusses this issue as well, mentioning several interesting possibilities. As the most commonly used methods he names the maximum and the average function but which combination function is best remains an open question. We do not intend to answer this question, but rather to contribute to the debate. In our opinion, from the point of view of ensembles, using the maximum of scores has some decisive disadvantages whereas the average does seem to make more sense.

To understand the disadvantage of using the maximum score, consider Figure 2, depicting the result of the maximum function (blue X) as combination of the individual score vectors (red crosses), in comparison to the average in Figure 1. The maximum as a combination function results in the upper bound of all individual scores and, hence, has a tendency to overestimate the outlierness. This also means that a single result that is far off, overestimating the scores for some objects, will determine the ensemble result (e.g., Figure 2(b)). Errors for different objects, contributed by different individual outlier score vectors, can lead the maximum combination actually more off while all individual outlier scores are not too bad overall (e.g., Figure 2(a)) — remember that this is a toy example for the rankings of two objects only. For a realistic scenario with $n \gg 2$ objects, it would be even more likely that some individual score vector is off for some single object (i.e., in a one-dimensional subspace) and all the other score vectors would not matter at all for this object.

This contradicts the very intuition of building ensembles. Errors of each individual score vector for single objects are strongly emphasized, and an error of a single ensemble member assigning a high outlier score to some object cannot be compensated for, even if all other detectors would be correct. This drawback counteracts one of the most fundamental benefits that one can expect from using an ensemble method: the correction of errors committed by single ensemble members. Let us note that keeping the maximum scores from different outlier models, if these models are learned in different subspaces, could be an approach to the problem of "multiview" outliers that we mentioned in the introduction — and that is somehow orthogonal to the ensemble or consensus idea.

On the other hand, using the average of scores has been theoretically advocated [75] for the combination of outlier scores based on (local) density estimates (as used by many classic methods such as LOF and its variants [9; 56; 38; 39] or the kNN outlier model and its variants [60; 4; 74]).[1] Building the average of different density estimates allows to abstract from the individual errors of these density estimates and, instead, to reason about the *expected* error. This reasoning

---

[1]Note that the set of methods using *local* density estimates is not restricted to the so-called *local outlier detection* methods. These are two different notions of "local", as elaborated by Schubert et al. [65].
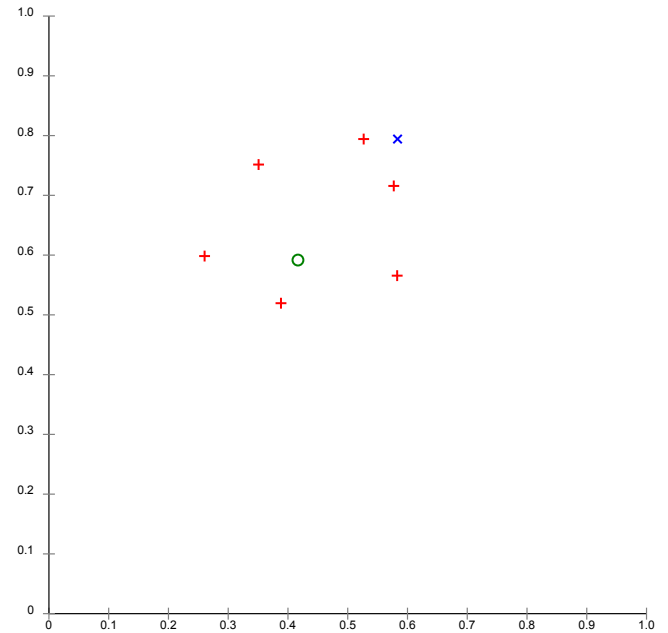
(a) Diverse results and their maximum
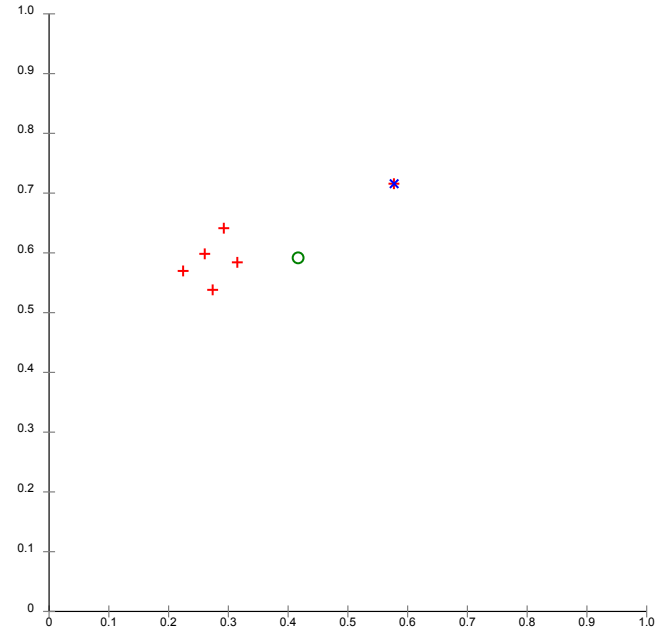
(b) Clustered results and their maximum

Figure 2: The maximum for score combination results in an upper bound of the result vectors.

might open up possibilities to improve our theoretic understanding of the benefit of ensembles for outlier detection.

However, the choice of a particular combination function will also remain application dependent. If the cost of missing a single outlier is much higher than the cost for a high false alarm rate, using the maximum combination is certainly worth considering. On the other hand, in an application scenario where the cost for false negatives is very high but missing some outliers might not hurt too much, maybe even the minimum as a combination method for score methods

may be a good choice. This would mean that all individual methods would have to assign a high outlier score to an object in order to actually count this object as an outlier. If just one of the ensemble members assigns a small outlier score, the minimum ensemble would use this smallest score for this object.

5.3 Combination of Rankings

Combining rankings provided by outlier methods, ignoring the actual outlier scores, could be seen as a particular way of normalization. But there is a considerable amount of literature in databases and in information retrieval on the combination of rankings that opens up possibilities for transferring known results from these areas to the particular problem of outlier detection. The feature bagging method [45], for example, was discussed in combination with a breadth-first traversal rank combination, i.e., taking the top ranked object from each individual ranking, then the second rank and so on. This is almost equivalent to using the maximum as score combination function[2] and, thus, has the same pros and cons.

Most methods, however, use the outlier scores and not only the rankings. This might be motivated by the assumption that the scores and their relative differences have at least *some* meaning (an assumption that actually might be debatable, in particular for high dimensional data due to an effect analogous to the concentration of distances [76]). See also the discussion of normalization issues by Aggarwal [2, Section 4.1].

5.4 Challenges

As challenges for future research on the aspect of combining several outlier detection results to a consensus or ensemble ranking or score vector, we see the following issues and questions for research:

- Improved normalizations of outlier scores via a better understanding of score distributions can possibly improve interpretability of individual scores and, as a consequence, can lead to a smoother combination of outlier scores to an ensemble.

- Can scores actually be converted into "outlier probabilities" [23; 40]? How would the success of such a conversion be evaluated (the problem of "calibration")?

- Should the calibration of outlier scores depend on the application? With different costs of false negatives or false positives, maybe the outlier scores should also get a bias in the more important direction. This question might have parallels in cost sensitive learning and the application of problem dependent loss functions.

- What are the effects of different combination functions? Which combination function is suitable for which application scenario?

- How to transfer rank accumulation procedures known in different areas such as databases and information retrieval to outlier ranking with its particular requirements?

- Can we improve our theoretical understanding of why unsupervised ensembles actually work?

6. CONCLUSIONS

Aggarwal [2] discussed algorithmic patterns, identified traces of the ensemble idea in the literature, and touched upon more or less likely options for future transfer of concepts from supervised ensembles for classification to ensembles for unsupervised outlier detection. Complementing his overview, we focused on the fundamental ingredients for success in building ensembles for unsupervised outlier detection. These are (1) learning accurate but (2) diverse models and (3) combining these models (or a selection thereof). For all these aspects, the literature provides not more than some first steps and insights which we sketched in this paper. As we point out, there are many opportunities to improve, for all aspects, we listed some challenges and issues for future research. It is our hope to stimulate research on the surprisingly neglected but very interesting and promising topic of ensembles for outlier detection.

7. REFERENCES

[1] N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA*, pages 504–509, 2006.

[2] C. C. Aggarwal. Outlier ensembles [position paper]. *ACM SIGKDD Explorations*, 14(2):49–58, 2012.

[3] C. C. Aggarwal. *Outlier Analysis*. Springer, 2013.

[4] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discoverys (PKDD), Helsinki, Finland*, pages 15–26, 2002.

[5] H. Ayad and M. Kamel. Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors. In *4th International Workshop on Multiple Classifier Systems (MCS), Guildford, UK*, pages 166–175, 2003.

[6] J. Azimi and X. Fern. Adaptive cluster ensemble selection. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI), Pasadena, CA*, pages 992–997, 2009.

[7] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[8] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley&Sons, 3rd edition, 1994.

[9] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 93–104, 2000.

[10] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6:5–20, 2005.

---

[2]Different from the maximum score combination, this breadth-first rank combination introduces a discretization and the resulting ranking depends on the order of the individual rankings for the traversal.

[11] R. Caruana, M. Elhaway, N. Nguyen, and C. Smith. Meta clustering. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China*, pages 107–118, 2006.

[12] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):Article 15, 1–58, 2009.

[13] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):823–839, 2012.

[14] X. H. Dang, I. Assent, R. T. Ng, A. Zimek, and E. Schubert. Discriminative features for identifying and interpreting outliers. In *Proceedings of the 30th International Conference on Data Engineering (ICDE), Chicago, IL*, 2014.

[15] X. H. Dang, B. Micenkova, I. Assent, and R. Ng. Outlier detection with space transformation and spectral analysis. In *Proceedings of the 13th SIAM International Conference on Data Mining (SDM), Austin, TX*, pages 225–233, 2013.

[16] T. G. Dietterich. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems (MCS), Cagliari, Italy*, pages 1–15, 2000.

[17] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the 13th International Conference on Machine Learning (ICML), Bari, Italy*, pages 105–112, 1996.

[18] A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. Systematic construction of anomaly detection benchmarks from real data. In *Workshop on Outlier Detection and Description, held in conjunction with the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA*, 2013.

[19] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning (ICML), Washington, DC*, pages 186–193, 2003.

[20] X. Z. Fern and W. Lin. Cluster ensemble selection. *Statistical Analysis and Data Mining*, 1(3):128–141, 2008.

[21] A. L. N. Fred and A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[22] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek. On using class-labels in evaluation of clusterings. In *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC*, 2010.

[23] J. Gao and P.-N. Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China*, pages 212–221, 2006.

[24] J. Ghosh and A. Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315, 2011.

[25] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[26] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[27] S. T. Hadjitodorov and L. I. Kuncheva. Selecting diversifying heuristics for cluster ensembles. In *7th International Workshop on Multiple Classifier Systems (MCS), Prague, Czech Republic*, pages 200–209, 2007.

[28] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3):264–275, 2006.

[29] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[30] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[31] M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan. Unifying dependent clustering and disparate clustering for non-homogeneous data. In *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC*, pages 593–602, 2010.

[32] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[33] N. Iam-On and T. Boongoen. Comparative study of matrix refinement approaches for ensemble clustering. *Machine Learning*, 2013.

[34] W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore*, pages 577–593, 2006.

[35] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analyis*. John Wiley&Sons, 1990.

[36] F. Keller, E. Müller, and K. Böhm. HiCS: high contrast subspaces for density-based outlier ranking. In *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC*, 2012.

[37] E. M. Knorr and R. T. Ng. A unified notion of outliers: Properties and computation. In *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, CA*, pages 219–222, 1997.

[38] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, pages 1649–1652, 2009.

[39] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, pages 831–838, 2009.

[40] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, pages 13–24, 2011.

[41] H.-P. Kriegel, P. Kröger, and A. Zimek. Subspace clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4):351–364, 2012.

[42] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, pages 444–452, 2008.

[43] L. I. Kuncheva and S. T. Hadjitodorov. Using diversity in cluster ensembles. In *Proceedings of the 2004 IEEE International Conference on Systems, Man, and Cybernetics (ICSMC), The Hague, Netherlands*, pages 1214–1219, 2004.

[44] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.

[45] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, pages 157–166, 2005.

[46] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3:1–39, 2012.

[47] M. J. A. N. C. Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale, Paris, 1785.

[48] M. Meila. Comparing clusterings – an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning (ICML), Bonn, Germany*, pages 577–584, 2005.

[49] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander. Density-based clustering validation. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, 2014.

[50] E. Müller, S. Günnemann, I. Färber, and T. Seidl. Discovering multiple clustering solutions: Grouping objects in different views of the data. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia*, page 1220, 2010.

[51] E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm. Outlier ranking via subspace analysis in multiple views of the data. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, pages 529–538, 2012.

[52] E. Müller, M. Schiffer, and T. Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, pages 434–445, 2011.

[53] M. C. Naldi, A. C. P. L. F. Carvalho, and R. J. G. B. Campello. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, 27(2):259–289, 2013.

[54] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan*, pages 368–383, 2010.

[55] N. Nguyen and R. Caruana. Consensus clusterings. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha, NE*, pages 607–612, 2007.

[56] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In *Proceedings of the 19th International Conference on Data Engineering (ICDE), Bangalore, India*, pages 315–326, 2003.

[57] D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems (KAIS)*, 19(3):361–394, 2009.

[58] N. Pham and R. Pagh. A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China*, 2012.

[59] Z. J. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Paris, France*, pages 717–726, 2009.

[60] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 427–438, 2000.

[61] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[62] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39, 2010.

[63] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA*, pages 1047–1058, 2012.

[64] E. Schubert, A. Zimek, and H.-P. Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, 2014.

[65] E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014.

[66] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong. A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery*, 26(2):332–397, 2013.

[67] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[68] A. Topchy, A. Jain, and W. Punch. Clustering ensembles: Models of concensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

[69] A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. Fred. Analysis of consensus partition in cluster ensemble. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), Brighton, UK*, pages 225–232, 2004.

[70] G. Valentini and F. Masulli. Ensembles of learning machines. In *Proceedings of the 13th Italian Workshop on Neural Nets, Vietri, Italy*, pages 3–22, 2002.

[71] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235, 2010.

[72] L. Vendramin, P. A. Jaskowiak, and R. J. G. B. Campello. On the combination of relative clustering validity criteria. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), Baltimore, MD*, pages 4:1–12, 2013.

[73] J. Yang, N. Zhong, Y. Yao, and J. Wang. Local peculiarity factor and its application in outlier detection. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, pages 776–784, 2008.

[74] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, pages 813–822, 2009.

[75] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander. Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, 2013.

[76] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012.

[77] A. Zimek and J. Vreeken. The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning*, 2013.