2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

AORS: Affinity-based Outlier Ranking Score

Shaohong Zhang, Hau-San Wong, Wen-Jun Shen, and Dongqing Xie

Abstract— **Outlier ranking methods can provide a quantitative measure to evaluate the outlierness of data instances in data clustering and attract great interest in pattern recognition and data mining communities. However, it has been pointed out that the diverse scaling ranges of these scores bring difficulty to result interpretation. Moreover, popular outlier ranking scores based on simple distance measures might not accurately reflect the complex affinity among data points. In this paper, we propose a new outlier ranking method based on consensus affinity of a cluster ensemble. Two new outlier ranking scores generalized from well-known clustering evaluation measures, Rvv from the RAND measure and ARIvv from Adjusted Rand Index (ARI), are adopted for outlierness evaluation. Compared to other outlierness ranking measures, the two new measures have the desired bounds without additional transformations. Consistent with the improvement of Adjusted Rand Index (ARI) over RAND, we find that ARIvv also significantly outperforms Rvv. Benefiting from the consensus affinity of a cluster ensemble, our proposed method with the ARIvv score provides significant improvement beyond a number of competing algorithms on public UCI benchmark data sets. Studies with both theoretical analysis and experimental validation show the effectiveness of our proposed methods.**

I. Introduction

Clustering techniques have been commonly used to discover knowledge in a lot of practical applications. However, due to the lack of supervisory information, clustering results generally suffer degradations from outliers and noise. The most related work on these kinds of applications is outlier ranking, whose objective is to identify an outlierness score for each data point to evaluate their potential to be an outlier, i.e., to be inconsistent with the other points [1]. The traditional unsupervised outlier ranking methods are statistics-based, which are obtained through some statistical characteristics of data [2]–[4], or based on fitting the data to different distributions [5], [6]. There are also the distance-based ranking methods [7], which consider the distances between certain important data points and a fraction of all other points [8], [9]. Variations to this class of methods use alternative distances in the K Nearest Neighbor (KNN) step [10], [11]. Another important class outlier ranking methods is density-based [12]–[14], which is inspired from the density-based clustering algorithms. The basic algorithm within this category, Local outlier factor (LOF) [12], focuses on the comparison between the local density of data instances with those of their neighbors. Although various related outlier ranking methods have been proposed, applying these methods to clustering is still an open problem. Specifically, widely used outlier ranking scores based on simple distance measures might not accurately reflect the complex affinity among data points and clusters. It is difficult to select a uniform distance/similarity measure and set of parameters for different types of data. In addition, existing outlier ranking scores are not easily interpretable in the context of clustering, and they might differ greatly in their scales and ranges [7], [15].

In view of the mentioned difficulties above, in this paper, we propose a new outlier ranking method based on consensus affinity of cluster ensembles. Here, affinity means that a pair of data points is within the same cluster. Interestingly, affinity is believed to be closely related to clustering. More specifically, affinity propagation [16], which iteratively exchanges similarity information between data points to identify representative exemplars for clusters, represents one of the more recent clustering approaches. In addition, a number of well-known pair-counting clustering evaluation measures [17]–[19] can be also regarded as a specific function of affinity. These measures are usually calculated based on the extent to which the cluster and class memberships of pairs of data points agree or disagree. Generalized pair-counting measures based on affinity also begin to attract great attention in recent years [19]–[21]. For these measures, affinity could be an important source of information when the true labels of the data points are unavailable.

We use the consensus affinity of cluster ensembles as the referenced knowledge for outlier ranking in unsupervised clustering. To our best knowledge, there are no related reports on this line of work. Cluster ensemble combines multiple individual clustering solutions into a consensus one to improve performance over that of any single clustering algorithm [22]–[24]. In general, there are two phases in a cluster ensemble approach: (i) to combine multiple individual solutions into a consensus one, usually in the form of a consensus matrix [21], [22], [25], a hypergraph [22], or a bipartite graph [23], [24]; (ii) to obtain a final partition from the consensus structure generated in the first phase using clustering algorithms [25], graph partition algorithms [22]–[24] or other methods. However, although quite a number of cluster ensemble methods have been proposed, it is still difficult to select a suitable cluster ensemble solution for different data sets with diverse characteristics. In addition, almost all of the cluster ensemble methods require prior knowledge of the number of clusters. Unfortunately, this kind of knowledge

Shaohong Zhang and Dongqing Xie are with the Department of Computer Science, Guangzhou University, Guangzhou, P.R. China (email: zimzsh@gmail.com); Hau-San Wong and Wen-Jun Shen are with the Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong, P.R. China (email: cshswong@cityu.edu.hk)

The work was partially supported by a grant from National Natural Science Foundation of China [No. 61202273], a grant from Natural Science Foundation of Guangdong Province [No. S2012040007206], a grant from Department of Education in Guangdong province [No. 2013KJCX0144], a grant from Guangzhou Education Bureau Science Foundation for Yangcheng Scholars (Project No. 10A033D), and a grant from the City University of Hong Kong [No. 7004047].

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

is usually unavailable, which makes application of cluster ensemble methods difficult. As a result, it is important to take into consideration these two problems when we make use of a cluster ensemble. Motivated by our previous studies on the generalized Adjusted Rand Index between similarity matrices [20], [21], we propose a new outlier ranking method based on generalized clustering evaluation measures, without the need to solve the cluster ensemble problem and to have the prior knowledge of the number of clusters. The effectiveness of our new methods and the interpretability of our results are investigated in detail, from both the perspectives of theoretical analysis and experimental validation.

Therefore, the most important contribution in this study is the introduction of an affinity-based outlier ranking method for clustering. To our best knowledge, there are no related studies based on consensus affinity of cluster ensembles. As the concept of affinity is closely related to clustering and cluster evaluation, it is hopeful that good performance can be achieved if we choose suitable outlier ranking scores. The second contribution is our adoption of consensus affinity without either the need to select a particular cluster ensemble method, or the prior knowledge of the number of clusters, which alleviates the main problems that affect previous methods and makes our approach more readily applicable to different types of data sets. Another advantage of our outlier ranking scores is their bounded ranges and their interpretability, which are important when comparing their performance with those of other similar competitors.

II. The proposed framework

As mentioned above, the affinity-based outlier ranking process includes two main phases: (i) Consensus structure generation: to compute a consensus structure from a cluster ensemble, and (ii) score evaluation, to rank each individual data instance using the affinity-based outlier ranking scores.

A. Consensus structure generation

The ensemble formulation phases include two steps: (i) generation of multiple clustering solutions using different settings; and (ii) computation of the consensus structure from these clustering solutions.

Generation of different clustering solutions

Throughout this paper, we use Kmeans [26] as our clustering algorithm. Considering a data set of N instances, $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$, in which each data instance $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \cdots, x_{iD}\}$ has D features. In an unsupervised manner, Kmeans assigns each data instance to one of K clusters and forms a partition P to minimize the total distances from data instances to their cluster centroid.

An important problem is how to select proper parameters for these individual clustering solutions, so as to generate a consensus structure which is as informative as possible. This problem is still an active topic in cluster ensemble research [27], [28]. In previous studies, individual clustering solutions are usually constructed using different numbers of clusters or different feature subsets. In this paper, we combine these two methods into one: for each independent clustering trial, we randomly select the number of clusters in the range of $[\sqrt{N}, 2\sqrt{N}]$, and the number of features from $[D/2, D]$, where N is the number of data instances, and D is the number of features. Note that our approach is not dependent on prior knowledge of the number of clusters.

Generation of the consensus structure

The second step generates a consensus structure from different individual clustering solutions constructed above. Note that there are different possible kinds of consensus structures derived from a cluster ensemble, such as a consensus matrix [20], [21], [25], [29], a bipartite graph [23], [24], or a hypergraph [22]. We choose to use the consensus matrix generated from individual clustering solutions as the reference for our affinity-based outlier ranking scores. Specifically, for each clustering solution $\{P^{(t)}\}_{t=1}^{T}$ generated by Kmeans in the first step, we transform it to the corresponding $N \times N$ co-association matrix as follows

$$M_{ij}^{(t)} = \begin{cases} 1 & \text{if } \exists\ k,\ \boldsymbol{x}_i \in P_k^{(t)} \text{ and } \boldsymbol{x}_j \in P_k^{(t)} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We can construct the consensus matrix of these individual clustering solutions from their co-association matrices using

$$\mathcal{M} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{M}^{(t)} \tag{2}$$

B. Affinity-based Outlier Ranking Score (AORS)

How to design a suitable outlier ranking score based on the consensus matrix is the most important problem for this paper. Intuitively, the entries in the i-th row of the consensus matrix, i.e., $[\mathcal{M}_{i1}, \mathcal{M}_{i2}, \cdots, \mathcal{M}_{iN}]$, can be viewed as an affinity measure between the data instance $\boldsymbol{x}_i$ and all the other instances. If $\mathcal{M}_{ij}$ is close to one(or zero), the two data instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have a large probability to be inside the same cluster (or in two different clusters). In other words, the degree of affinity uncertainty for these two points is small in this case. On the other hand, when the degree of affinity uncertainty becomes larger (i.e., the affinity of the point $\boldsymbol{x}_i$ with the other points is only vaguely known), a desirable measure should result in a significantly different value from that in the former case. Interesting, preliminary studies on the overall degree of affinity uncertainty from the consensus matrix has been investigated in our recent work [21] based on a generalized formulation of the well-known Adjusted Rand Index (ARI) measure [18]. In this paper, we shall design the Affinity-based Outlier Ranking Score (AORS) in a similar spirit as in [21]. In the following, we shall first provide a brief introduction of the RAND measure [17] and the Adjusted Rand Index (ARI) measure [18]. We then propose the new measures to evaluate the Affinity-based Outlier Ranking Score (AORS) for each individual data instance.

C. The RAND and the Adjusted Rand Index (ARI) measures

RAND [17] is a well-known measure to compare the similarity between two partitions in a pair-counting manner. For two partitions P and Q on a data set X with N instances,

let $\delta()$ be the indicator function with $\delta(true) = 1$ and $\delta(false) = 0$, RAND between partitions P and Q can be represented as

$$RAND(P,Q) = \frac{a_0 + d_0}{\lambda_0} \tag{3}$$

where

$$\begin{aligned}
a_0 &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(\boldsymbol{x}_i) = P(\boldsymbol{x}_j)\ \&\ Q(\boldsymbol{x}_i) = Q(\boldsymbol{x}_j)) \\
b_0 &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(\boldsymbol{x}_i) = P(\boldsymbol{x}_j)\ \&\ Q(\boldsymbol{x}_i) \neq Q(\boldsymbol{x}_j)) \\
c_0 &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(\boldsymbol{x}_i) \neq P(\boldsymbol{x}_j)\ \&\ Q(\boldsymbol{x}_i) = Q(\boldsymbol{x}_j)) \\
d_0 &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(\boldsymbol{x}_i) \neq P(\boldsymbol{x}_j)\ \&\ Q(\boldsymbol{x}_i) \neq Q(\boldsymbol{x}_j)) \\
\lambda_0 &= a_0 + b_0 + c_0 + d_0
\end{aligned} \tag{4}$$

Note that the four factors can be interpreted in the following way: a_0 is the number of point pairs assigned to the same class in both partitions P and Q; b_0 is the number of point pairs assigned to the same class in partition P and to different classes in partition Q; c_0 is the number of point pairs assigned to different classes in partition P and to the same class in partition Q; d_0 is the number of point pairs assigned to different classes in both partitions P and Q.

RAND is well-known to be affected by the number of clusters K, and tends to result in inflated scores when K is large [18], [29]–[31]. An improved measure, Adjusted Rand Index (ARI) [18], is proposed to alleviate the dependence on the number of clusters. ARI can also be represented in a pair-counting manner as follows:

$$ARI(P,Q) = \frac{a_0 - \frac{(a_0+b_0)(a_0+c_0)}{\lambda_0}}{\frac{1}{2}(a_0 + b_0 + a_0 + c_0) - \frac{(a_0+b_0)(a_0+c_0)}{\lambda_0}} \tag{5}$$

D. The proposed measures

As mentioned above, we can extend RAND and ARI in a similar spirit as in [21], to evaluate the consistency between two data points. Specifically, for two data instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and the consensus matrix $\mathcal{M}$, denote $\boldsymbol{u} = [u_1, u_2, \cdots, u_N] = [\mathcal{M}_{i1}, \mathcal{M}_{i2}, \cdots, \mathcal{M}_{iN}]$ and $\boldsymbol{v} = [v_1, v_2, \cdots, v_N] = [\mathcal{M}_{j1}, \mathcal{M}_{j2}, \cdots, \mathcal{M}_{jN}]$, we can compute a number of new factors as follows

$$\begin{aligned}
a &= \sum_{i=1}^{N} u_i v_i, b = \sum_{i=1}^{N} u_i(1 - v_i) \\
c &= \sum_{i=1}^{N} (1 - u_i)v_i, d = \sum_{i=1}^{N} (1 - u_i)(1 - v_i) \\
\lambda &= a + b + c + d
\end{aligned} \tag{6}$$

We now propose two new measures, which are expressed in terms of the above factors as follows:

$$Rvv(\boldsymbol{u}, \boldsymbol{v}) = \frac{a + d}{\lambda} \tag{7}$$

$$ARIvv(\boldsymbol{u}, \boldsymbol{v}) = \frac{a - \frac{(a+b)(a+c)}{\lambda}}{0.5(a + b + a + c) - \frac{(a+b)(a+c)}{\lambda}} \tag{8}$$

Compared to the original clustering evaluation measures (3) and (5), the difference of these two new measures comes from the factors a, b, c and d, which are based on only one row of the consensus matrix $\mathcal{M}$ for each data point, e.g., $\boldsymbol{u} = [\mathcal{M}_{i1}, \cdots, \mathcal{M}_{iN}]$ for $\boldsymbol{x}_i$ and $\boldsymbol{v} = [\mathcal{M}_{j1}, \mathcal{M}_{j2}, \cdots, \mathcal{M}_{jN}]$ for $\boldsymbol{x}_j$. These two proposed measures, $Rvv(\boldsymbol{u}, \boldsymbol{v})$ and $ARIvv(\boldsymbol{u}, \boldsymbol{v})$, can be used to evaluate the consistency of the two corresponding data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in terms of the extent of their agreements in clustering assignments, similar to the cases in our recent work [21]. However, in this paper, we focus on estimating the outlierness of the data points. Intuitively, $Rvv(\boldsymbol{u}, \boldsymbol{u})$ and $ARIvv(\boldsymbol{u}, \boldsymbol{u})$, can be regarded as the evaluation of the extent of agreement of clustering assignments between the data point $\boldsymbol{x}$ and itself. Similar to the original measures in cluster evaluation, data points with larger $Rvv(\boldsymbol{u}, \boldsymbol{u})$ values or $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ values can be regarded as more likely to be non-outliers. On the other hand, data points with small score values are likely to be outliers. Details of the interpretation and proofs of these two new scores are presented in the following section.

The complete Affinity-based Outlier Ranking Score (AORS) approach is summarized in Algorithm 1.

Algorithm 1: AORS

input : $N \times D$ matrix of data set $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$;
input : number of individual partitions T;
output: data instance importance score list $\boldsymbol{s}$.

$K_{max} \leftarrow \lceil 2\sqrt{N} \rceil$;
for *each individual clustering solution* $t \leftarrow 1$ **to** T **do**
  sample a feature size $D^{(t)}$ from $\{D/2, \cdots, D\}$;
  generate a reduced subset $X^{(t)}$ of $D^{(t)}$ features sampled at random;
  sample a cluster number $K^{(t)}$ from 2 to K_{max};
  cluster $X^{(t)}$ with `kmeans`;
  compute co-association matrix $\boldsymbol{M}^{(t)}$ using (1);
end
compute the consensus matrix $\mathcal{M}$ using (2);
select a similarity measure (ARIvv or Rvv);
for *each row in the consensus matrix* $\boldsymbol{u} = [\mathcal{M}_{i1}, \cdots, \mathcal{M}_{iN}]$ **do**
  compute the outlier ranking:
  $\boldsymbol{s}_i = ARIvv(\boldsymbol{u}, \boldsymbol{u})$ using (8) or
  $\boldsymbol{s}_i = Rvv(\boldsymbol{u}, \boldsymbol{u})$ using (7) ;
end
return outlier ranking score list $\boldsymbol{s}$;

III. Interpretation of proposed scores

During these few years, the interpretation of outlier scores begins to attract great interests from leading research groups [4], [7], [15]. Specifically, the ranges of the output score values of many outlier detection methods may differ greatly among different data sets or even among different categories within a single data set. As a result, a consistent range for outlier scores is regarded as important, and various methods are proposed to transform different outlier scores to a unified range $[0, 1]$ for better interpretation [7]. On the other hand, our proposed affinity-based outlier ranking scores, ARIvv and Rvv, are already in this range. We shall discuss these in detail in this section with the related proofs.

Proposition 1: For a data point $\boldsymbol{x}$ associated with the i-th row in the consensus matrix $\boldsymbol{u} = [\mathcal{M}_{i1}, \cdots, \mathcal{M}_{iN}]$, the affinity-based outlier ranking score $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ has a bounded range of [0, 1].

Proof: For easier expression, we first use a simpler notation for the measure ARIvv as follows

$$\begin{aligned} &A = a, B = 0.5(a + b + a + c) \\ &C = \frac{(a+b)(a+c)}{\lambda}, ARIvv(\boldsymbol{u}, \boldsymbol{v}) = \frac{A - C}{B - C} \end{aligned} \tag{9}$$

For the numerator of the affinity-based outlier ranking score $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ in Eq. (8), we can obtain

$$\begin{aligned} &A - C = a - \frac{(a+b)(a+c)}{\lambda} \\ =&\frac{a\lambda - (a+b)(a+c)}{\lambda} = \frac{a(a+b+c+d) - (a+b)(a+c)}{\lambda} \\ =&\frac{(a^2 + ab + ac + ad) - (a^2 + ab + ac + bc)}{\lambda} = \frac{ad - bc}{\lambda} \end{aligned} \tag{10}$$

As $\lambda > 0$, we can focus on the numerator as follows:

$$\begin{aligned} &ad - bc \\ =&\sum_{i=1}^{N} u_i u_i \sum_{i=1}^{N} (1-u_i)(1-u_i) - \sum_{i=1}^{N} u_i(1-u_i) \sum_{i=1}^{N} (1-u_i)(u_i) \\ =&\sum_{i=1}^{N} {u_i}^2 \sum_{i=1}^{N} (1-u_i)^2 - (\sum_{i=1}^{N} u_i(1-u_i))^2 \end{aligned} \tag{11}$$

We can obtain $ad - bc \geq 0$ from the Cauchy-Schwarz inequality $|\langle \boldsymbol{u}, \boldsymbol{v} \rangle|^2 \leq \langle \boldsymbol{u}, \boldsymbol{u} \rangle \langle \boldsymbol{v}, \boldsymbol{v} \rangle$, and therefore the numerator of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ is not less than zero.

On the other hand, we can easily see that

$$\begin{aligned} B - A =& 0.5(a + b + a + c) - a = 0.5(b + c) \\ =& 0.5(b + b) = b \geq 0 \end{aligned} \tag{12}$$

The third step uses the fact that $b = c$ for $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ between $\boldsymbol{u}$ and itself. Therefore, for the factors A, B, and C defined in Eq. (9), we can obtain

$$C \leq A \leq B \tag{13}$$

Thus, we can obtain

$$0 \leq ARIvv(\boldsymbol{u}, \boldsymbol{u}) = \frac{A - C}{B - C} \leq 1 \tag{14}$$

The condition for the lower bound of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ can be derived directly from Eq. (11) as follows

$$u_i = (1 - u_i), \ \forall i \Rightarrow u_i = 0.5, \ \forall i \tag{15}$$

which means that under this condition the assignment of the point $\boldsymbol{x}$ to a cluster becomes the most difficult, considering that the probabilities for $\boldsymbol{x}$ and all the other points to be in the same cluster or different clusters are all 0.5, i.e., all random). In this case, $\boldsymbol{x}$ has the maximum uncertainty in cluster assignment.

The upper bound of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$ can be derived directly from Eq. (12) and Eq. (20) as follows

$$ARIvv(\boldsymbol{u}, \boldsymbol{u}) = \frac{A - C}{B - C} = \frac{A - C}{B - C} = \frac{A - C}{A + b - C} \tag{16}$$

Thus, the maximum condition turns out to be

$$b = \sum_{i=1}^{N} u_i(1 - u_i) = 0 \tag{17}$$

As $0 \leq u_i \leq 1$, we can obtain the maximum condition as

$$u_i = \{0, 1\}, \ \forall i \tag{18}$$

which means that the point $\boldsymbol{x}$ is certain to be in the same cluster or in different clusters with all the other points in the cluster ensemble, i.e., $\boldsymbol{x}$ has the minimum uncertainty in cluster assignment.

Thus, the affinity-based outlier ranking score for the data point $\boldsymbol{x}$, $ARIvv(\boldsymbol{u}, \boldsymbol{u})$, has a bounded range of [0, 1]. ■

Proposition 2: The affinity-based outlier ranking score $Rvv(\boldsymbol{u}, \boldsymbol{u})$ has a bounded range of [0.5, 1].

Proof: As in the case of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$, we have $b = c$. Thus

$$Rvv(\boldsymbol{u}, \boldsymbol{u}) = \frac{a + d}{\lambda} = \frac{a + d}{a + b + c + d} = \frac{a + d}{a + 2b + d} \tag{19}$$

Since $a \geq 0, b \geq 0, d \geq 0$, we have

$$0 < Rvv(\boldsymbol{u}, \boldsymbol{u}) = \frac{a + d}{a + 2b + d} \leq 1 \tag{20}$$

As in the case of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$, the maximum condition is

$$u_i = \{0, 1\}, \ \forall i \tag{21}$$

Since a and d are dependent, the minimum case is different from that in the case of $ARIvv(\boldsymbol{u}, \boldsymbol{u})$. Note that the denominator in the measure $Rvv(\boldsymbol{x}, \boldsymbol{x})$ is a constant

$$\lambda = a + b + c + d = N \tag{22}$$

Thus, the minimum value of $Rvv(\boldsymbol{u}, \boldsymbol{u})$ is dependent on its numerator

$$\begin{aligned} &a + d = \sum_{i=1}^{N} u_i u_i + \sum_{i=1}^{N} (1-u_i)(1-u_i) \\ =&\sum_{i=1}^{N} u_i^2 + \sum_{i=1}^{N} (1 - 2u_i + u_i^2) = \sum_{i=1}^{N} (2u_i^2 - 2u_i + 1) \\ =&\sum_{i=1}^{N} 2(u_i - 0.5)^2 + 0.5N \geq 0.5N \end{aligned} \tag{23}$$

We can also observe that the minimum condition is the same as that in $ARIvv(\boldsymbol{u}, \boldsymbol{u})$

$$u_i = 0.5, \ \forall i \tag{24}$$

Thus,

$$Rvv(\boldsymbol{u}, \boldsymbol{u}) = \frac{a+b}{\lambda} \geq \frac{0.5N}{N} = 0.5 \tag{25}$$

i.e., the minimum value of $Rvv(\boldsymbol{u}, \boldsymbol{u})$ is 0.5, which is different from that in $ARIvv(\boldsymbol{u}, \boldsymbol{u})$.

Thus, we have proved that the affinity-based outlier ranking $scoreRvv(\boldsymbol{u}, \boldsymbol{u})$ has a bounded range of [0.5, 1]. ■

IV. Experiments

A. Experiment setup

In this section, we verify the effectiveness of AROS using a number of public data sets. We begin by describing the data sets we have used, the parameter selection process, the list of previous algorithms for comparison with our algorithm, and the evaluation metrics.

Data sets: We use a number of well-known benchmark datasets from the UCI machine learning repository[1], including UCI-BCW, UCI-Chart, UCI-HandWritingDigit, UCI-Pendigits, UCI-Vertebral, and UCI-Wine.

Parameter selection: Kmeans with the Euclidean distance is used as the basic clustering algorithm. There is only one parameter to be considered: T, the number of different clustering solutions in (2). We set $T = 100$ as in related cluster ensemble studies [29], [32]. We also investigate the dependence of the results on different T values in the experiments.

Competing algorithms: The following outlier ranking methods are used for the comparison in this study:(1) Random removal. A certain proportion of data points are selected at random and excluded from clustering. Statistical evaluation results of this method should not differ too much and they can serve as the baseline references for the other methods; (2) Outlier ranking methods based on K Nearest Neighbors (KNN) [10]. The number of nearest neighbors is set to 10 as in [10]; (3) Outlier ranking methods based on the aggregates of K Nearest Neighbors (KnnAgg) [11]. The number of nearest neighbors is also set to 10. (4) Local outlier factor (LOF) [12], which uses a range of MinPts values rather than a specific number of nearest neighbors. The lower bound of MinPts is also set to 10 as suggested in the original paper [12], and the upper bound of MinPts is set to 40 (In the original paper [12], no explicit selection of the upper bound of MinPts is provided, and different data sets are discussed with MinPts values from 35 to 45). (5) Our affinity-based outlier ranking methods with the Rvv score (Rvv). (6) Our affinity-based outlier ranking methods with the ARIvv score (ARIvv).

Evaluation metrics: Three popular measures, i.e., RAND [17], Adjusted Rand Index (ARI) [18], and Normalized Mutual Information (NMI) [22] are widely adopted to evaluate clustering performance. We use all these three measures for the comparison between our methods with the other ones, to avoid the possible bias of any single measure. For further comparison between the two proposed methods, i.e., Rvv and ARIvv, only NMI is used as the evaluation measure, in order to remove the possible dependence between RAND and Rvv and that between ARI and ARIvv.

For all the experiments, the results are based on averaging across 50 trials if not otherwise specified.

[1] http://archive.ics.uci.edu/ml/

B. Comparison of AORS with competing algorithms

We compare AORS with a number of competing algorithms on six UCI data sets, with different proportions of removed samples. Specifically, the removed proportions of samples range from 1% to 10%, and we evaluate the Kmeans clustering performance on the remaining data. Results on the three different measures, RAND [17], Adjusted Rand Index (ARI) [18], and Normalized Mutual Information (NMI) [22], are shown in Figures 1, 2, and 3, respectively. Although they have similar formulations, ARIvv significantly outperforms Rvv in all the data sets based on the three different measures. The RAND measure is believed to have less discrimination power when the number of clusters increases [18], [21], [30], [31], which motivates the adoption of an alternative measure such as ARI. Interestingly, ARIvv shows similar advantage when compared to Rvv in these experiments. From the figures, we can see that ARIvv has better performance in most cases when compared to the other competitors. Also, we can find that the performance of ARIvv tends to increase as the portion of removed data increases, which suggests that the removed samples have an adverse effect on the clustering.

C. Comparison of ARIvv and Rvv scores

We have further investigated the correlation between ARIvv and Rvv in Figure 4. For each sample in each trial, its ARIvv score is used as the x-axis value and its Rvv score as the y-axis one. The overall correlation of ARIvv and Rvv for all the samples in each data set is shown above each scatter plot. Interestingly, we can observe that different data sets could result in very different distributions of the (ARIvv, Rvv) points. The largest correlation value is 0.8457 while the smallest is 0.1981. This suggests that the ARIvv scores do not have a strong correlation with the Rvv scores. In addition, we can observe that the Vertebral data set has a number of samples appearing in the top left, which indicates that these samples have large Rvv scores but small ARIvv ones. Also, we can find ARIvv have a wider distribution than Rvv, which is consistent with the well-recognized understanding that the ARI measure value has greater discrimination power than the RAND measure in clustering performance evaluation [18], [21], [31].

D. Performance studies of ARIvv with different ensemble sizes.

Note that our AORS method with ARIvv score has only one parameter, i.e., the ensemble size T. Thus, it is interesting to study the effect of this parameter on the resulting performance. We use different ensemble sizes from 50 to 500

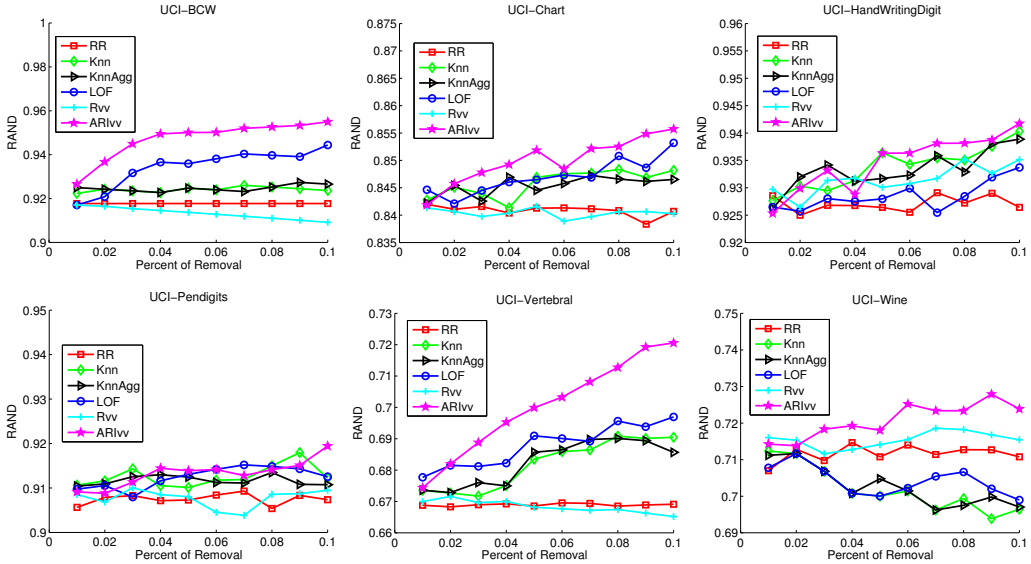

Fig. 1. Performance on UCI data sets: RAND. ARIvv outperforms all the other competitors.

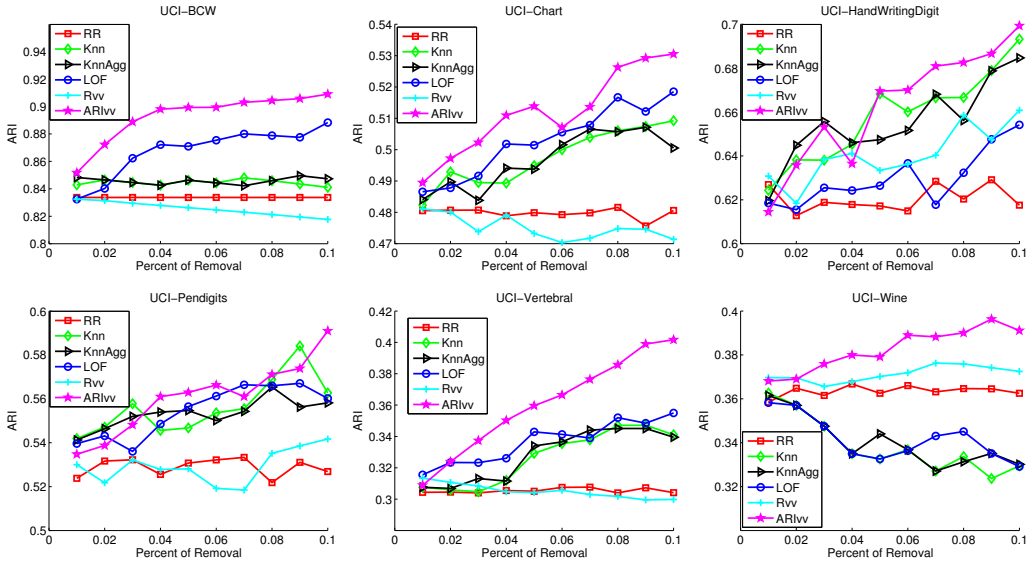

Fig. 2. Performance on UCI data sets: ARI. ARIvv outperforms all the other competitors.

with a step value of 50. Clustering performances of each data set under different numbers of ensembles are shown in Figure 5. In the figure, each column corresponds to one ensemble size. The colors of the columns vary smoothly from cyan to magenta, in accordance with the results from the worst to the best. We can also observe that the performance variation is quite small with respect to different ensemble sizes. Also, we can find that the standard deviation, indicated by the error bar at the top of each column, is very small when compared to the column height. These observations suggest that our method is quite robust and not affected too much by the value of T.

V. Conclusion

It is important to identify and exclude outliers in data clustering. However, existing outlier ranking methods are generally based on simple distance measures tailored for particular data sets, and cannot accurately reflect the complex affinity among data points for different data types. In this paper, we propose a new affinity-based outlier ranking method with two new outlier ranking scores, in a similar spirit of well-known generalized clustering evaluation measures. Our methods are based on the affinity between data points and thus can be universally applicable to different data sets of diverse characteristics. To our best knowledge, our work is the first effort in making effective use of the consensus affinity of a cluster ensemble to achieve this objective. We also propose to generalize well-known pair-counting clus-

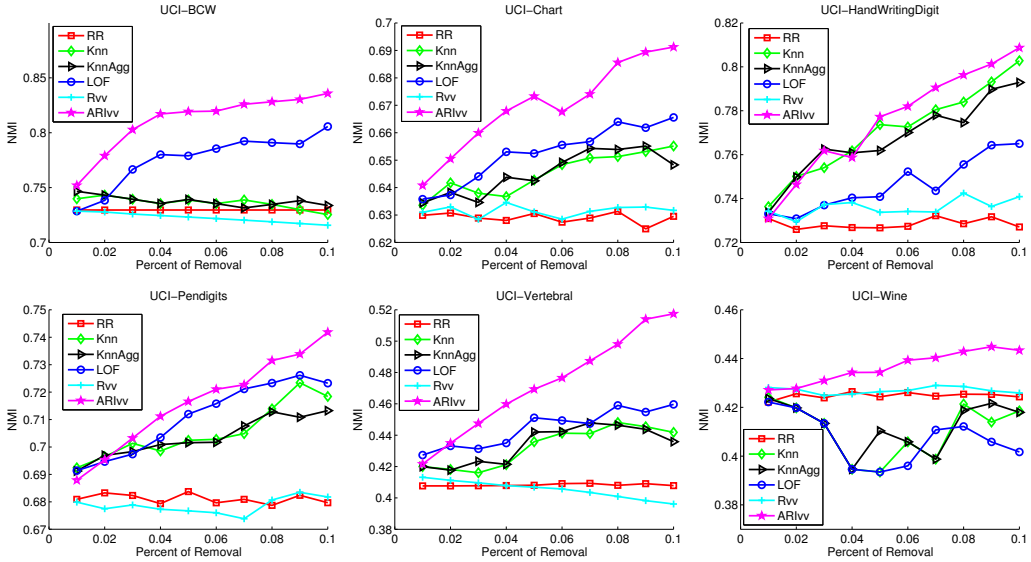

Fig. 3. Performance on UCI data sets: NMI. ARIvv outperforms all the other competitors.

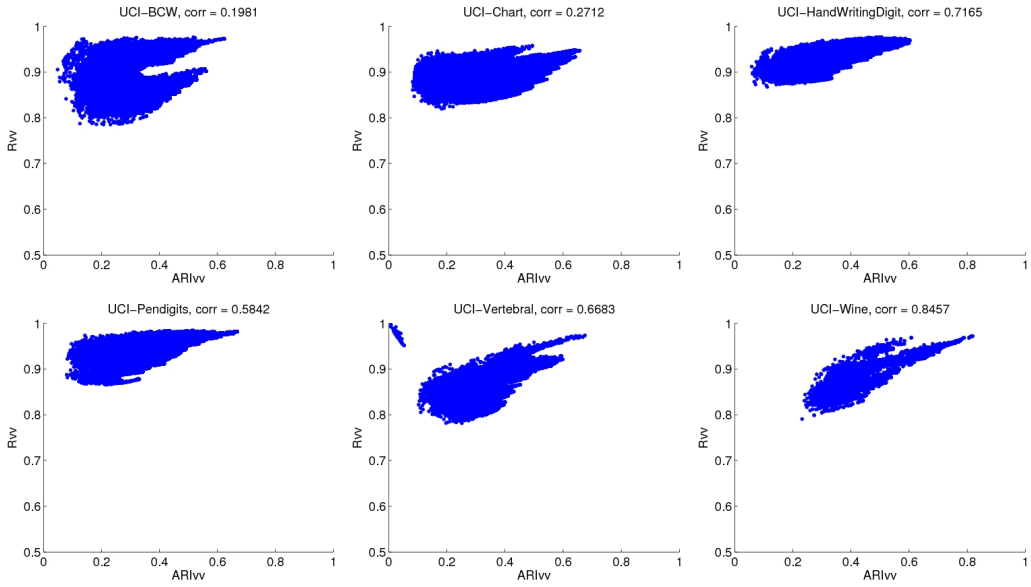

Fig. 4. Comparison of ARIvv and Rvv scores. ARIvv has a wider range than Rvv.

tering measures, including the Rand Index measure and the Adjusted Rand Index (ARI) measure, to compute the outlier ranking scores. This approach alleviates the main problems of traditional clustering ensembles by avoiding the need to obtain a final consensus solution and the prior knowledge of the number of clusters. In addition, our proposed outlier ranking scores have bounded ranges and straightforward interpretation, which are verified through theoretical analysis. Experiments on a number of UCI benchmark data sets show that our method based on the adjusted rand index (ARIvv) significantly outperforms competing algorithms.

In future, we would like to study the generalized forms of other clustering evaluation measures. We have demonstrated the importance of selecting suitable outlier ranking scores in this paper, and it will be interesting to study which measures can be used to further improve the performance of our methods. In addition, it is also important to investigate how we can adapt the various measures for different data sets with diverse characteristics.

REFERENCES

[1] V. Barnett and T. Lewis, *Outliers in statistical data (3rd edition.)*. John Wiley and Sons, 1994.

[2] M. Stephens, "Tests based on edf statistics," *Goodness-of-fit Techniques*, vol. 68, pp. 97–193, 1986.

[3] R. Bremer, "Outliers in statistical data," *Technometrics*, vol. 37, no. 1, pp. 117–118, 1995.

[4] P. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

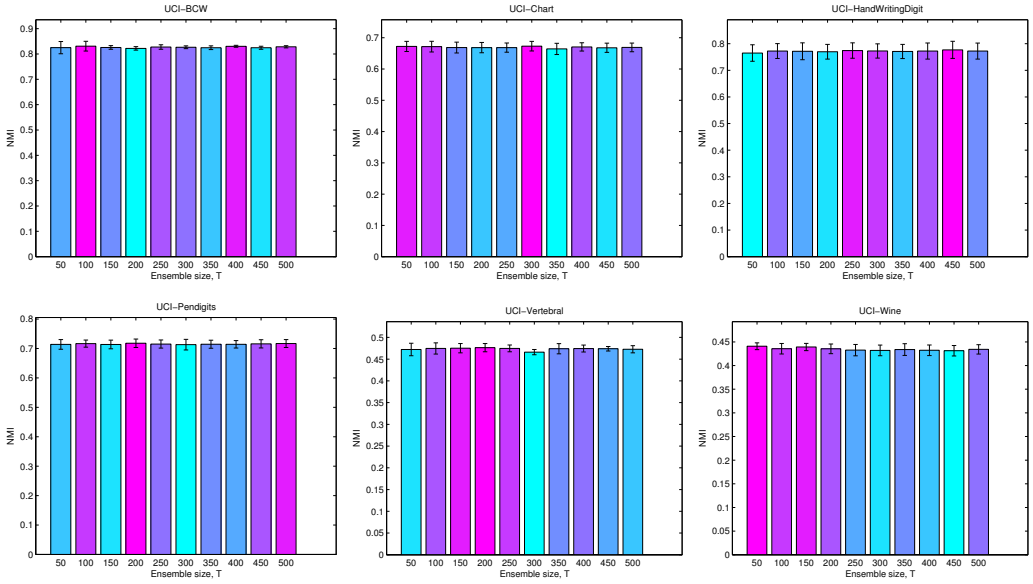

Fig. 5. Performance on UCI data sets with different ensemble sizes. Results are shown in colors varying smoothly from cyan to magenta, in accordance with the results from the worst to the best. The difference and the standard deviation are small, which shows that our ARIvv method is rather insensitive to the parameter T.

[5] P. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, pp. 212–223, 1999.

[6] J. Hardin and D. Rocke, "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator," *Computational Statistics & Data Analysis*, vol. 44, no. 4, pp. 625–638, 2004.

[7] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores." in *2011 SIAM International Conference on Data Mining*, 2011, pp. 13–24.

[8] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the International Conference on Very Large DataBases*. Citeseer, 1998, pp. 392–403.

[9] E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.

[10] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *ACM SIGMOD Record*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

[11] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces." Springer, 2002, pp. 43–78.

[12] M. Breunig, H. Kriegel, R. Ng, J. Sander *et al.*, "Lof: identifying density-based local outliers," *Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.

[13] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," *Advances in Knowledge Discovery and Data Mining*, pp. 813–822, 2009.

[14] T. de Vries, S. Chawla, and M. Houle, "Finding local anomalies in very high dimensional space," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 128–137.

[15] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, "On evaluation of outlier rankings and outlier scores," in *2012 SIAM International Conference on Data Mining*.

[16] B. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[17] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[18] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[19] R. J. G. B. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment." *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[20] S. Zhang and H.-S. Wong, "Arimp: A generalized adjusted rand index for cluster ensembles," in *the 20th International Conference on Pattern Recognition (ICPR 2010)*, Istanbul, Turkey., 2010.

[21] S. Zhang, H.-S. Wong, and Y. Shen, "Generalized adjusted rand indices for cluster ensembles," *Pattern Recognition*, vol. 45, no. 6, pp. 2214–2226, 2012.

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

[23] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proceedings of the twenty-first International Conference on Machine learning*, 2004.

[24] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 99, pp. 2396–2409, 2011.

[25] A. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, 2005.

[26] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[27] L. Kuncheva, S. Hadjitodorov, and L. Todorova, "Experimental comparison of cluster ensemble methods," in *Proceedings of the 9th International Conference on Information Fusion*. IEEE, 2006, pp. 1–7.

[28] J. Azimi and X. Fern, "Adaptive cluster ensemble selection." in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, 2009, pp. 992–997.

[29] S. Zhang, H. Wong, Y. Shen, and D. Xie, "A new unsupervised feature ranking method for gene expression data based on consensus affinity," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 9, no. 4, pp. 1257–1263, 2012.

[30] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information." in *Advances in Neural Information Processing Systems 15*, 2003, pp. 505–512.

[31] N. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA, 2009.

[32] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, "Consensus unsupervised feature ranking from multiple views," *Pattern Recognition Letters*, vol. 29, no. 5, pp. 595–602, 2008.