Unsupervised Ensemble Learning for Mining Top-n Outliers*

Jun Gao[1], Weiming Hu[1], Zhongfei(Mark) Zhang[2], and Ou Wu[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{jgao,wmhu,wuou}@nlpr.ia.ac.cn

[2] Dept. of Computer Science, State Univ. of New York at Binghamton,
Binghamton, NY 13902, USA
zhongfei@cs.binghamton.edu

Abstract. Outlier detection is an important and attractive problem in knowledge discovery in large datasets. Instead of detecting an object as an outlier, we study detecting the n most outstanding outliers, i.e. the top-n outlier detection. Further, we consider the problem of combining the top-n outlier lists from various individual detection methods. A general framework of ensemble learning in the top-n outlier detection is proposed based on the rank aggregation techniques. A score-based aggregation approach with the normalization method of outlier scores and an order-based aggregation approach based on the distance-based Mallows model are proposed to accommodate various scales and characteristics of outlier scores from different detection methods. Extensive experiments on several real datasets demonstrate that the proposed approaches always deliver a stable and effective performance independent of different datasets in a good scalability in comparison with the state-of-the-art literature.

1 Introduction

Outlier detection is an important knowledge discovery problem in finding unusual events and exceptional cases from large datasets in many applications such as stock market analysis, intrusion detection, and medical diagnostics. Over the past several decades, the research on outlier detection varies from the global computation to the local analysis, and the descriptions of outliers vary from binary interpretations to probabilistic representations. Global outlier detection [3,4,5] identifies an observational object with a binary label by the global computation. Local outlier detection [6,7,8,9] provides a probabilistic likelihood called outlier score to capture how likely an object is considered as an outlier. Outlier scores can be used not only to discriminate outliers from normal data, but also to rank all the data in a database, such as the top-n outlier detection. There are other efforts that transform the unsupervised outlier detection to a classification via artificially generated outliers [10].

* This work is supported in part by NSFC(Grant No. 60825204, 60935002 and 60903147), NBRPC(2012CB316400) and US NSF(IIS-0812114, CCF-1017828) as well as the Zhejiang University – Alibaba Financial Joint Lab.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 418–430, 2012.
© Springer-Verlag Berlin Heidelberg 2012

Although there are numerous outlier detection methods proposed in the literature, no one method performs better than the others under all circumstances, and the best method for a particular dataset may not be known a priori. Each detection method is proposed based on the specific priori knowledge. For example, the nearest neighbor based methods assume that the feature space is well enough to discriminate outliers from normal data, while the classification based and the statistical methods need to suppose the distributions of outliers and normal objects, respectively. Hence, their detection performances vary with the nature of data. This setting motivates a fundamental information retrieval problem - the necessity of an ensemble learning of different detection methods to overcome their drawbacks and to increase the generalization ability, which is similar to *meta-search* that aggregates query results from different search engines into a more accurate ranking. Like *meta-search*, ensemble learning in the top-n outlier detection is more valuable than the fusion of the binary labels, especially in large databases. There is the literature on the ensemble learning of outlier detection, such as [13,14,15]. However, all these efforts state the problem of effectively detecting outliers in the sub-feature spaces. Since the work of Lazarevic and others focuses on the fusion of the sub-feature spaces, these methods are very demanding in requiring the full spectrum of outlier scores in the datasets that prevents them from the fusion of the top-n outlier lists in many real-world applications.

Although the problem of ensemble learning in the top-n outlier detection shares a certain similarity to that of *meta-search*, they have two fundamental differences. First, the top-n outlier lists from various individual detection methods include the order information and outlier scores of n most outstanding objects. Different detection methods generate outlier scores in different scales. This requires the ensemble framework to provide a unified definition of outlier scores to accommodate the heterogeneity of different methods. Second, the order-based rank aggregation methods, such as Mallows Model [18], can only combine the information of the order lists with the same length, which prevents the application of these rank aggregation methods in the fusion of top-k outlier lists. Because, for a particular dataset, there are always several top-k outlier lists with various length used to measure the performance and effectiveness of a basic outlier detection method. In order to address these issues, we propose a general framework of ensemble learning in the top-n outlier detection shown in Figure 1, and develop two fusion methods: the score-based aggregation method (*SAG*) and the order-based aggregation method (*OAG*). To the best of our knowledge, this is the first attempt to the ensemble learning in the top-n outlier detection. Specifically, the contributions of this paper are as follows:

- We propose a score-based aggregation method (*SAG*) to combine the top-n outlier lists given by different detection methods without supervision. Besides, we propose a novel method for transforming outlier scores to posterior probabilities, which is used to normalize the heterogeneous outlier scores.
- We propose an order-based aggregation method (*OAG*) based on the distanced-based Mallows model [16] to aggregate the different top-n outlier lists without supervision, which can deal with the fusion of top-k outlier lists with various length. This method only adopts the order information, which avoids the normalization of outlier scores.

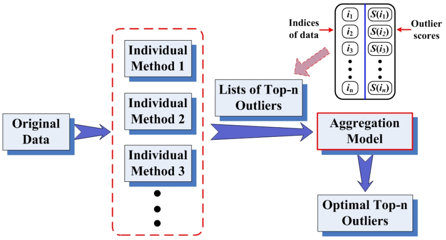


Fig. 1. The general framework of ensemble learning

- Extensive experiments on real datasets validate the effectiveness of these aggregation methods, where several state-of-the-art outlier detection methods, including the nearest neighbor based and the classification based methods, are selected as the individual methods for the ensemble learning. Besides, the robustness of the proposed aggregation methods is evaluated based on the Uniform noise and the Gaussian noise.

The remainder of this paper is organized as follows. Section 2 introduces the framework of ensemble learning in the top-n outlier detection and the two novel aggregation methods: the score-based and the order-based methods. Section 3 reports the experimental results. Finally, Section 4 concludes the paper.

2 Methodologies

We first introduce the general framework and the basic notions of ensemble learning in the top-n outlier detection, and then introduce the score-based method with a unified outlier score and the order-based method based on the distance-based Mallows model, respectively.

2.1 Framework and Notions of Ensemble Learning

Let $X = [x_1, x_2, x_3, \ldots, x_d]$ be an object in a dataset D, where d is the number of attributes and $|D|$ is the number of all the objects.

As shown in Figure 1, there are K individual detection methods that process the original data in parallel. Essentially, all the individual methods return outlier scores rather than binary labels to generate the top-n outlier lists, where the number n is determined by users. The top-n outlier list σ_i assigned to the i-th individual method is represented as $(\sigma^{-1}(1), S(i_1); \cdots; \sigma^{-1}(n), S(i_n))$, where $\sigma^{-1}(i)$ denotes the index of the object assigned to rank i and $S(\sigma^{-1}(i))$ is its outlier score. Correspondingly, $\sigma(i)$ is the rank assigned to object X_i. Let R_n be the set of all the top-n orderings over $|D|$ objects, and $d : R_n \times R_n \longrightarrow \mathbb{R}$ be the distance between two top-n lists, which should be a right-invariant metric. This means that the value of $d(\pi, \sigma)|\forall \pi, \sigma \in R_n$ does not depend on how objects are indexed.

The aggregation model combines K orderings $\{\sigma_i\}_{i=1}^{K}$ to obtain the optimal top-n outlier list. Clearly, the literature with respect to the fusion of sub-feature spaces [13,14,15] can be included in this framework by using the detection model in a special sub-feature space as an individual method. In this paper, we only focus on the unsupervised aggregation models based on the order information and outlier scores.

2.2 Score-Based Aggregation Approach (SAG)

Since a top-n outlier list σ_i contains the order information and the corresponding outlier scores, it is straightforward that combining these outlier scores from different methods improves the detection performance. As mentioned in the previous section, outlier scores of the existing methods have different scales. For example, outlier scores vary from zero to infinity for the nearest based method [6], while lying in the interval $[-1, 1]$ for the classification based method [10]. In this subsection, an effective method is proposed to transform outlier scores to posterior probability estimates. Compared with outlier scores, the posterior probability based on Bayes' theorem provides a robust estimate to the information fusion and a spontaneous measure of the uncertainty in outlier prediction. Without loss of generality, we assume that the higher $S(i)$, the more probable X_i to be considered as an outlier. Let Y_i be the label of X_i, where $Y_i = 1$ indicates that X_i is an outlier and $Y_i = 0$ if X_i is normal. According to Bayes' theorem,

$$P(Y_i = 1|S(i)) = \frac{P(S(i)|Y_i = 1)P(Y_i = 1)}{\sum_{l=0}^{1} P(S(i)|Y_i = l)P(Y_i = l)} = \frac{1}{1 + \frac{P(S(i)|Y_i=0)P(Y_i=0)}{P(S(i)|Y_i=1)P(Y_i=1)}} \quad (1)$$

Let $\varphi(i) = \frac{P(S(i)|Y_i=0)P(Y_i=0)}{P(S(i)|Y_i=1)P(Y_i=1)}$. $ln\big(\varphi(i)\big)$ can be considered as the discriminant function that classifies X_i as normal or outlier. Hence, $ln\big(\varphi(i)\big)$ can be simplified to a linear function, proportional to the Z-Score of $S(i)$ as follows:

$$\varphi(i) = exp\bigg(-\frac{S(i) - \mu}{std} + \tau \bigg) \quad (2)$$

where μ and std are the mean value and standard deviation of the original outlier scores, respectively. In large datasets, these statistics can be computed by sampling the original data. As a discriminant function, $ln\big(\varphi(i)\big) < 0$ means $(S(i) - \mu)/std > \tau$; the object X_i can be assigned as an outlier. In all the experiments, the default value of τ equals 1.5 based on Lemma 1.

Lemma 1: *For any distribution of outlier score $S(i)$, it holds that*

$$P\bigg(\frac{S(i) - \mu}{std} > \tau\bigg) \le \frac{1}{\tau^2}$$

Proof: *According to Chebyshev's inequality, it holds that,*

$$P\bigg(\frac{S(i) - \mu}{std} > \tau\bigg) \le P\bigg(|S(i) - \mu| > \tau \cdot std\bigg) \le \frac{std^2}{(\tau \cdot std)^2} = \frac{1}{\tau^2}$$

Lemma 1 shows a loose bound of deviation probability regardless of the distribution of outlier scores. Supposing that outlier scores follow a normal distribution, $\tau = 1.5$ means that much less than 10% of the objects deviate from the majority of data, which follows the definition of Hawkins outlier [1].

For a top-n outlier list σ_i, objects in the dataset may not be ranked by σ_i. The simple average posterior probabilities are not appropriate to the top-n ranking aggregation. Clearly, objects that appear in all the ranking lists should be more probable to be outliers than ones that are only ranked by a single list. Hence, we apply the following fusion rules which are proposed by Fox and Show [12].

$$rel(i) = n_d^r \sum_j rel_j(i) \quad r \in (-1, 0, 1) \tag{3}$$

where n_d is the number of the orderings that contain object X_i and $rel_j(i)$ is the normalized outlier score of X_i by the j-th individual method. When $r = 1$, the ultimate outlier score is composed of the number of the orderings n_d and the sum of its outlier scores. When $r = 0$, the result is only the sum of its outlier scores. When $r = -1$, it is equivalent to the average outlier scores of the orderings containing X_i. According to Eq. 1 and Eq. 2, the posterior probabilities can be used to normalize outlier scores directly. The detailed steps of *SAG* are shown in Algorithm 1.

Algorithm 1. Score-based aggregation method (*SAG*)

```
Input: $\psi = \{\sigma_k\}_{k=1}^{K}, \gamma$
1. Transform outlier scores in $\psi$ to posterior probabilities according to Eq. $\{1\ 2\}$.
2. Construct an union item pool $U$ including all objects in $\psi$, and denote the size of $U$ as $|U|$.
3. Compute the normalized outlier score $\{rel(i)\}_{i=1}^{|U|}$ for each object in $U$ according to Eq. 3.
4. Sort objects in $U$ based on the normalized outlier scores, and output the optimal list $\pi$.
Output: $\pi$
```

2.3 Order-Based Aggregation Approach (OAG)

Given a judge ordering σ and its expertise indicator parameter θ, the Mallows model [16]generates an ordering π given by the judge according to the formula:

$$P(\pi|\theta, \sigma) = \frac{1}{Z(\sigma, \theta)} exp(\theta \cdot d(\pi, \sigma)) \tag{4}$$

where

$$Z(\sigma, \theta) = \sum_{\pi \in R_n} exp(\theta \cdot d(\pi, \sigma)) \tag{5}$$

According to the right invariance of the distance function, the normalizing constant $Z(\sigma, \theta)$ is independent of σ, which means $Z(\sigma, \theta) = Z(\theta)$. The parameter θ is a non-positive quantity and the smaller the value of θ, the more concentrated at σ the ordering π. When θ equals 0, the distribution is uniform meaning that the ordering given by the judge is independent of the truth.

An extended Mallows model is proposed in [17] as follows:

$$P(\pi|\boldsymbol{\theta},\boldsymbol{\sigma}) = \frac{1}{Z(\boldsymbol{\sigma},\boldsymbol{\theta})} P(\pi) exp\Big(\sum_{i=1}^{K} \theta_i \cdot d(\pi,\sigma_i)\Big) \tag{6}$$

where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_K) \in R_n^K$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \in \mathbb{R}^K$, $P(\pi)$ is a prior, and the normalizing constant

$$Z(\boldsymbol{\sigma},\boldsymbol{\theta}) = Z(\boldsymbol{\theta}) = \sum_{\pi \in R_n} P(\pi) exp\Big(\sum_{i=1}^{K} \theta_i \cdot d(\pi,\sigma_i)\Big) \tag{7}$$

In this extended model, each ordering σ_i is returned by a judge for a particular set of objects. θ_i represents the expertise degree of the i-th judge. Eq. 6 computes the probability that the true ordering is π, given the orderings $\boldsymbol{\sigma}$ from K judges and the degrees of their expertise.

Based on the hypothesis of the distance-based Mallow model, we propose a generative model of *OAG*, which can be described as follows:

$$P(\pi,\boldsymbol{\sigma}|\boldsymbol{\theta}) = P(\boldsymbol{\sigma}|\boldsymbol{\theta},\pi)P(\pi|\boldsymbol{\theta}) = P(\pi)\prod_{i=1}^{K} P(\sigma_i|\theta_i,\pi) \tag{8}$$

The true list π is sampled from the prior distribution $P(\pi)$ and σ_i is drawn from the Mallows model $P(\sigma_i|\theta_i,\pi)$ independently. For the ensemble learning of top-n outlier lists, the observed objects are the top-n outlier lists $\boldsymbol{\sigma}$ from various individual detection methods, and the unknown object is the true top-n outlier list π. The value of the free parameter θ_i depends on the detection performance of the i-th individual method. The goal is to find the optimal ranking π and the corresponding free parameter θ_i which maximize the posteriori probability shown in Eq. 6. In this work, we propose a novel EM algorithm to solve this problem. For obtaining an accurate estimation of θ_i by the EM-based algorithm, we construct the observed objects by applying several queries with different lengths $\{N_q\}_{q=1}^{Q}$, where $N_1 = n$ and $N_{q/1} > n$. Clearly, it is to compute the parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ by considering the information of different scales. In this paper, the default value of Q is 4 and the lengths meet the following requirement: $N_q = q \cdot n$.

2.4 Inference and Algorithm for OAG

The EM algorithm is widely used for finding the maximum likelihood estimates in the presence of missing data. The procedure includes two steps. First, the expected value of the complete data log-likelihood with respect to the unobserved objects $\phi = \{\pi_q | \pi_q \in R_{N_q}\}_{q=1}^{Q}$, the observed objects $\psi = \{\boldsymbol{\sigma_q} | \boldsymbol{\sigma_q} \in R_{N_q}^K\}_{q=1}^{Q}$, and the current parameter estimate $\boldsymbol{\theta'} = (\theta'_1, \cdots, \theta'_K)$. Second, compute the optimal parameter $\boldsymbol{\theta}$ that maximizes the expectation value in the first procedure. According to the Mallows model and the extended Mallows model, we have the following Lemmas:

Lemma 2: *The expected log-likelihood $\zeta(\boldsymbol{\theta},\boldsymbol{\theta'})$ meets the following formula*

$$\zeta(\boldsymbol{\theta},\boldsymbol{\theta'}) = \quad E[\log P(\phi,\psi|\boldsymbol{\theta})|\psi,\boldsymbol{\theta'}] = \sum\nolimits_{(\pi_1,\cdots,\pi_Q)} L(\boldsymbol{\theta}) \cdot U(\boldsymbol{\theta'}) \tag{9}$$

where

$$L(\boldsymbol{\theta}) = \sum_{q=1}^{Q} \log P(\pi_q) - \sum_{q=1}^{Q}\sum_{i=1}^{K} \log Z_q(\theta_i) + \sum_{q=1}^{Q}\sum_{i=1}^{K} \theta_i \cdot d(\pi_q, \sigma_q^i) \tag{10}$$

$$U(\boldsymbol{\theta'}) = \prod_{q=1}^{Q} P(\pi_q | \boldsymbol{\theta'}, \boldsymbol{\sigma_q}) \tag{11}$$

Lemma 3: *The parameter $\boldsymbol{\theta}$ maximizing the expected value $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta'})$ meets the following formula:*

$$\sum_{q=1}^{Q} E_{\theta_i}(d(\pi_q, \sigma_q^i)) = \sum_{(\pi_1, \cdots, \pi_Q)} \sum_{q=1}^{Q} d(\pi_q, \sigma_q^i) \cdot U(\boldsymbol{\theta'}) \tag{12}$$

The proofs for Lamma 2 and Lamma 3 are omitted due to lack of space. As shown in Lamma 3, the value of the right-hand side of Eq. 12 and the analytical expression of the left-hand side should be evaluated under the appropriate distance function to obtain the optimal $\boldsymbol{\theta}$. Before introducing the detailed procedure of our EM-based learning algorithm, we bring in an effective distance function $d(\pi, \sigma)$ between the top-n orderings π and σ, which is proposed in [18]. To keep this work self-contained, this distance function is introduced as follows.

Definition 1: *Let F_π and F_σ be the elements of π and σ respectively. $Z = F_\pi \cap F_\sigma$ with $|Z| = z$. $P = F_\pi \setminus Z$, and $S = F_\sigma \setminus Z$ (note that $|P| = |S| = n - z = r$). Define the augmented ranking $\tilde{\pi}$ as π augmented with the elements of S assigned the same index $n+1$. Clearly, $\tilde{\pi}^{-1}(n+1)$ is the set of elements at position $n+1$ ($\tilde{\sigma}$ is defined similarly). Then, $d(\pi, \sigma)$ is the minimum number of the adjacent transpositions needed to turn $\tilde{\pi}$ to $\tilde{\sigma}$ as follows, where $I(x) = 1$ if $x > 0$, and 0 otherwise.*

$$d(\pi, \sigma) = \sum_{\substack{i=1 \\ \tilde{\pi}^{-1}(i) \in Z}}^{n} V_i(\tilde{\pi}, \tilde{\sigma}) + \sum_{\substack{i=1 \\ \tilde{\pi}^{-1}(i) \notin Z}}^{n} U_i(\tilde{\pi}, \tilde{\sigma}) + \frac{r(r+1)}{2} \tag{13}$$

where

$$V_i(\tilde{\pi}, \tilde{\sigma}) = \sum_{\substack{j=i \\ \tilde{\pi}^{-1}(j) \in Z}}^{n} I(\tilde{\sigma}(\tilde{\pi}^{-1}(i)) - \tilde{\sigma}(\tilde{\pi}^{-1}(j))) + \sum_{j \in \tilde{\pi}^{-1}(n+1)} I(\tilde{\sigma}(\tilde{\pi}^{-1}(i)) - \tilde{\sigma}(j))$$

$$U_i(\tilde{\pi}, \tilde{\sigma}) = \sum_{\substack{j=i \\ \tilde{\pi}^{-1}(j) \in Z}}^{n} 1$$

In each iteration of the EM process, $\boldsymbol{\theta}$ is updated by solving Eq. 12. Based on Definition 1, $E_{\theta_i}(d(\pi_q, \sigma_q^i))$ is computed as follows:

$$E_{\theta_i}(d(\pi_q, \sigma_q^i)) = \frac{N_q e^{\theta_i}}{1 - e^{\theta_i}} - \sum_{j=r+1}^{N_q} \frac{j e^{j\theta_i}}{1 - e^{j\theta_i}} + \frac{r(r+1)}{2} - r(z+1)\frac{e^{\theta_i(z+1)}}{1 - e^{\theta_i(z+1)}}$$

This function is a monotonous function of the parameter θ_i. For estimating the right-hand side of Eq. 12, we adopt the Metropolis algorithm introduced in [2] to sample from Eq. 6. Suppose that the current list is π_t. A new list π_{t+1} is achieved by exchanging the objects i and j, which are randomly chosen from all the objects in π_t. Let $r = P(\pi_{t+1}|\boldsymbol{\theta}, \boldsymbol{\sigma})/P(\pi_t|\boldsymbol{\theta}, \boldsymbol{\sigma})$. If $r \geq 1$, π_{t+1} is accepted as the new list, otherwise π_{t+1} is accepted with the probability r. Then, $\boldsymbol{\theta}$ can be computed by the line search approach with the average z of the samples. The steps of *OAG* are shown in Algorithm 2.

Algorithm 2. Order-based aggregation method (*OAG*)

```
Input: $\psi = \{\boldsymbol{\sigma_q}\}_{q=1}^{Q}$ with $|\sigma_q^i| = N_q$, $\boldsymbol{\theta}^{(0)}$, $\varepsilon$, $t = 1, T$
  1. Construct the sampling sets $(\pi_i, \cdots, \pi_Q) \in R_n^Q$ by the Metropolis algorithm from Eq. 6.
  2. Compute the value of the right-hand side of Eq. 12.
  3. Adopt the line search approach to compute $\boldsymbol{\theta}^{(t+1)}$ based on Eq. 12
  4. If $t = T$, or $\sum_{i=1}^{K} |\theta_i^{t+1} - \theta_i^t| < \varepsilon$, return $\boldsymbol{\theta}^{(t+1)}$ and the optimal top-n outlier list $\pi$
     estimated by the sampling procedure; else $t = t + 1$, goto the step 1.
Output: $\boldsymbol{\theta}, \pi$
```

3 Experiments

We evaluate the aggregation performances of *SAG* and *OAG* methods using a number of real world datasets. We measure the robust capabilities of *SAG* and *OAG* methods to the random rankers, which are generated based on the Uniform distribution and the Gaussian distribution, respectively.

3.1 Aggregation on Real Data

In this subsection, we make use of several state-of-the-art methods, including *LOF* [6], *K-Distance* [3], *LOCI* [7], *Active Learning* [10], and *Random Forest* [11] as the individual methods to return the original top-n outliers lists. Since the performances of *LOF* and *K-Distance* depend on the parameter K that determines the scale of the neighborhood, we take the default value of K as 2.5% of the size of a real dataset. Both *LOF* and *LOCI* return outlier scores for each dataset based on the density estimation. However, *K-Distance* [3] only gives objects binary labels. Hence, according to the framework of *K-Distance*, we compute outlier scores as the distance between an object and its Kth nearest neighbor. *Active learning* and *Random Forest* both transform outlier detection to classification based on the artificial outliers generated according to the procedures proposed in [10]. These two methods both compute outlier scores by the majority voting of the weak classifiers or the individual decision trees.

The real datasets used in this section consist of the Mammography dataset, the Ann-thyroid dataset, the Shuttle dataset, and the Coil 2000 dataset, all of which can be downloaded from the UCI database except for the Mammography dataset.[1] Table 1

[1] Thank Professor Nitesh.V.Chawla for providing this dataset, whose email address is *nchawla@nd.edu*

Table 1. Documentations of the real data

Dataset		Mammography	Ann-thyroid	Shuttle-1	Shuttle-2	Shuttle-3	Coil-2000
Number of data	normal	10923	3178	11478	11478	11478	5474
	outlier	260	73	13	39	809	348
Proportion of outliers		2.32%	2.25%	0.11%	0.34%	6.58%	5.98%

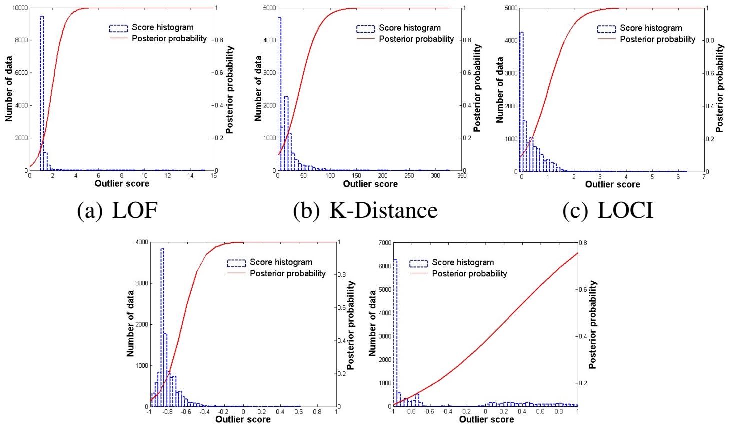
summarizes the documentations of these real datasets. All the comparing outlier detection methods are evaluated using *precision* and *recall* in the top-n outlier list σ as follows

$$Precision = TN/AN \qquad Recall = TN/ON$$

where TN is the number of outliers in ordering σ, AN is the length of σ, and ON is the number of outliers in the dataset. For the quantity AN equals ON in this work, *precision* has the same value with *recall*. Hence, only *precision* is used to measure the performance of each compared method in this section. Clearly, if all the objects in σ are outliers, its *precision* and *recall* both achieve the maximum value 100%. The *Breadth-first* and *Cumulative Sum* methods proposed in *Feature Bagging* [13] are used as the baselines. For *Feature Bagging* does not introduce how to normalize heterogeneous outlier scores, the original outlier scores are processed by the typical normalization method: $S_{norm}(i) = \frac{S(i)-mean}{std}$, where $mean$ is the average score of all the objects and std is the standard deviation of outlier scores. Besides, *Cumulative Sum* requires that every object should be given an outlier score by every individual method. However, for the top-n outlier lists, some objects lying in the ordering σ_i may not be ranked by σ_j. This means that *Cumulative Sum* cannot be applied in the fusion of the top-n outlier lists. Hence, we replace the sum of all the outlier scores with the average of the outlier scores from the individual methods containing the corresponding object for *Cumulative Sum*. The *Mallows Model* [18] is also used as the baseline. As discussed in the previous section, for this algorithm can not combine the basic lists $\boldsymbol{\sigma}$ with various lengths to achieve the true list π, it needs to use all the datasets to compute the expertise indicator parameter θ.

Table 2 lists the experimental results of the individual methods and all the aggregation methods. Figure 2 shows the posterior probability curves based on *SAG* for the individual methods on the Mammography dataset. It is very clear that different detection methods have different scales of outlier scores and posterior probability computed by *SAG* is a monotonic increasing function of outlier scores. In the individual method pool, *LOF* achieves the best performance on the Mammography and the Shuttle-2 datasets, and K-Distance achieves the best performance on the Shuttle-1 dataset. *LOCI* detects the most outliers on the Coil 2000 dataset with *Active learning*. *Random Forest* is superior to the other methods on the Ann-thyroid and Shuttle-3 datasets. However, none of the outliers is detected by *Random Forest* on the Shuttle-1,2 datasets. The above results have verified the motivation that there is a need of ensemble learning in the top-n outlier detection.

From Table 2, we see that *SAG* with $r = 1$ and *SAG* with $r = 0$ achieve the similar performance on all the real datasets. Clearly, for the probability-based *SAG* method,



(a) LOF (b) K-Distance (c) LOCI

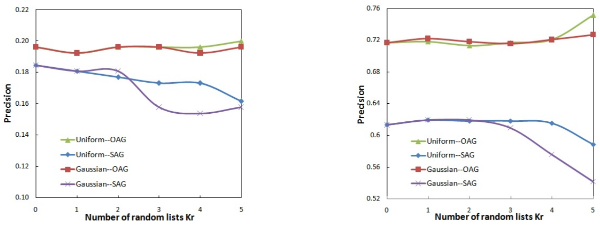
(d) Active Learning (e) Random Forest

Fig. 2. The posterior probability curves based on *SAG* and score histograms of various individual methods on the Mammography dataset

Table 2. The precisions in the top-n outlier lists for all the individual methods and the aggregation methods on the real data

Dataset **Method**	Mammography (Top 260)	Ann-thyroid (Top-73)	Shuttle-1 (Top-13)	Shuttle-2 (Top-39)	Shuttle-3 (Top-809)	Coil-2000 (Top-348)
LOF	19.0%	39.7%	23.1%	53.8%	28.4%	5.5%
K-Distance	13.8%	37.0%	29.8%	48.7%	34.5%	8.0%
LOCI	8.8%	28.8%	7.7%	33.3%	67.0%	8.9%
Active Learning	18.1%	28.8%	15.4%	0%	30.3%	8.9%
Random Forests	15.4%	41.1%	0%	0%	70.6%	8.6%
Average of All	15.0%	35.1%	15.2%	27.2%	46.2%	8.0%
Cumulative Sum	10.0%	31.5%	23.1%	**58.9%**	40.0%	10.3%
Breadth-first	14.2%	38.4%	0%	28.2%	46.9%	10.6%
Mallows Model	13.1%	38.4%	23.1%	51.3%	44.4%	8.0%
SAG (r= 1)	18.5%	34.2%	23.1%	48.7%	61.3%	9.8%
SAG (r= 0)	18.5%	34.2%	23.1%	48.7%	62.1%	9.5%
SAG (r=-1)	5.4%	26.0%	7.7%	43.6%	59.5%	**10.9%**
OAG	**19.7%**	**42.5%**	**30.8%**	53.8%	**71.7%**	9.1%

the number n_d of the individual top-n outlier lists contributes little to the final fusion performance. Compared with the above aggregation methods, the performance of *SAG* with $r = -1$ varies with the nature of the data dramatically. *SAG* with $r = -1$ achieves the best performance on the Coil 200 dataset. However, it performs more poorly than *SAG* with $r = \{1, 0\}$ and *OAG* on the other datasets. This demonstrates that the average of the unified outlier scores does not adapt to the fusion of the top-n lists. In general, since outlier scores are always either meaningless or inaccurate, the order-based aggregation method makes more sense than the score-based method. *OAG* achieves the



(a) Mammography (b) Shuttle-3

Fig. 3. The precisions of *OAG* and *SAG* $(r = 1)$ varying with the number of random lists K_r on the Mammography data and Shuttle-3 data

Table 3. The parameter $\boldsymbol{\theta}$ of all the individual methods and five random lists on the Mammography and Shuttle-3 datasets

Method Dataset		*LOF*	*K-Distance*	*LOCI*	*Active Learning*	*Random Forests*	*random lists (Average)*
Mammogrpahy	*Uniform-Noise*	-0.0058	-0.0039	-0.0058	-0.0052	-0.0039	-0.00014
	Gaussian-Noise	-0.0061	-0.0033	-0.0055	-0.0054	-0.0044	-0.00016
Shuttle-3	*Uniform-Noise*	-0.0014	-0.0016	-0.0014	-0.0018	-0.0037	-0.00001
	Gaussian-Noise	-0.0014	-0.0016	-0.0018	-0.0014	-0.0035	-0.00002

best performance than *SAG* on the Mammography, the Ann-thyroid, and the Shuttle-1,3 datasets. Both *Cumulative Sum* and *SAG* are score-based fusion methods. Table 2 shows that the performance of *SAG* is more stable and effective, especially *SAG* with $r = 1$. *Breath-first*, *Mallows Model*, and *OAG* are all the order-based fusion methods. Although *Breath-first* can be used in the aggregation of top-n outlier lists, it is sensitive to the order of the individual methods. *Mallows Model* supposes that there is a fixed expertise indicator parameter θ for an individual method regardless of the nature of the data. Experiment results indicates that this hypothesis is not appropriate for the ensemble learning in the top-n outlier detection. Overall, *SAG* and *OAG* both achieve the better performances than *Average of All* and the aggregation methods *Breadth-first*, *Cumulative Sum* and *Mallows Model*, which means that the proposed approaches deliver a stable and effective performance independent of different datasets in a good scalability.

3.2 Robustness of Two Aggregation Methods

In this subsection, the goal is to examine the behavior of the *SAG* and *OAG* methods when poor judges are introduced into the individual method pool. For a dataset D, the top-n outlier lists of the poor judges are generated from the underlying distribution U. First, the outlier scores of all the data are sampled from the distribution U. Then, the random top-n outlier lists are obtained by sorting all the data based on the outlier scores. In our experiments, two alternative definitions of U are used: Uniform distribution on the interval $[0, 1]$ and standard Gaussian distribution. The corresponding top-n lists are called *Uniform-Noise* and *Gaussian-Noise*. The individual method pool contains the previous five individual detection methods, and the K_r random lists of the poor judges, where K_r varies from 1 to 5.

For lack of the space, only the results on the Mammography dataset and the Shuttle-3 dataset are shown in the Figure 3. Clearly, *OAG* is more robust to the random poor judges than *SAG* regardless of *Uniform-Noise* or *Gaussian-Noise*. Especially, *OAG* achieves a better performance when the number K_r of random lists increases. Table 3 gives the value of the parameter $\boldsymbol{\theta}$ of the individual method pool on the Mammography and Shuttle-3 datasets. The parameter θ of each *Uniform-Noise* or *Gaussian-Noise* is close to zero. This demonstrates that *OAG* learns to discount the random top-n lists without supervision.

4 Conclusions

We have proposed the general framework of the ensemble learning in the top-n outlier detection in this paper. We have proposed the score-based method (*SAG*) with the normalized method of outlier scores, which is used to transform outlier scores to posterior probabilities. We have proposed the order-based method (*OAG*) based on the distance-based Mallows model to combine the order information of various individual top-n outlier lists. Theoretical analysis and empirical evaluations on several real data sets demonstrate that both *SAG* and *OAG* can effectively combine the state-of-the-art detection methods to deliver a stable and effective performance independent of different datasets in a good scalability, and *OAG* can discount the random top-n outlier lists without supervision.

References

1. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
2. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Journal of Biometrika 57(1), 97–109 (1970)
3. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. Journal of VLDB 8(3-4), 237–253 (2000)
4. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. Journal of ACM Computing Surveys (CSUR) 31(3), 264–323 (1999)
5. Barnett, V., Lewis, T.: Outliers in Statistic Data. John Wiley, New York (1994)
6. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. In: SIGMOD, pp. 93–104 (2000)
7. Papadimitriou, S., Kitagawa, H., Gibbons, P.: Loci: Fast outlier detection using the local correlation integral. In: ICDE, pp. 315–326 (2003)
8. Yang, J., Zhong, N., Yao, Y., Wang, J.: Local peculiarity factor and its application in outlier detection. In: KDD, pp. 776–784 (2008)
9. Gao, J., Hu, W., Zhang, Z(M.), Zhang, X., Wu, O.: RKOF: Robust Kernel-Based Local Outlier Detection. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS(LNAI), vol. 6635, pp. 270–283. Springer, Heidelberg (2011)
10. Abe, N., Zadrozny, B., Langford, J.: Outlier detection by active learning. In: KDD, pp. 504–509 (2006)
11. Breiman, L.: Random Forests. J. Machine Learning 45(1), 5–32 (2001)

12. Fox, E., Shaw, J.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), pp. 243–252 (1994)
13. Lazarevic, A., Kumar, V.: Feature bagging for outlier detection. In: KDD, pp. 157–166 (2005)
14. Gao, J., Tan, P.N.: Converting output scores from outlier detection algorithms into probability estimates. In: ICDM, pp. 212–221 (2006)
15. Nguyen, H., Ang, H., Gopalkrishnan, V.: Mining outliers with ensemble of heterogeneous detectors on random subspaces. Journal of DASFAA 1, 368–383 (2010)
16. Mallows, C.: Non-null ranking models. I. J. Biometrika 44(1/2), 114–130 (1957)
17. Lebanon, G., Lafferty, J.: Cranking: Combining rankings using conditional probability models on permutations. In: ICML, pp. 363–370 (2002)
18. Klementiev, A., Roth, D., Small, K.: Unsupervised rank aggregation with distance-based models. In: ICML, pp. 472–479 (2008)