Sequential Ensemble Method for Unsupervised Anomaly Detection

Huy Van Nguyen
Institute of Information Technology,
Vietnam Academy of Military Science & Technology
Email: vannguyenhuy.vn@gmail.com

Trung Thanh Nguyen
Faculty of IT,
Le Quy Don Technical University
Email: trungthanhnt@gmail.com

Quang Uy Nguyen
Faculty of IT,
Le Quy Don Technical University
Email: quanguyhn@gmail.com

Abstract—In data mining, anomaly detection aims at identifying the observations which do not conform to an expected behavior. To date, a large number of techniques for anomaly detection have been proposed and developed. Recently, researchers have paid their attention to ensemble methods to improve the accuracy of anomaly detection algorithms. Particularly, Sequential Ensemble Method (SEQ) proposed recently has shown significant improvement over other techniques. The idea of SEQ is to evaluate the scores of samples by using a second algorithm with respect to the first algorithm's output. In other words, an algorithm is firstly used to choose a set of the highest suspect abnormal samples (D_{ref}) and then a second algorithm is applied to evaluate the final score of each data samples in the dataset with respect to only D_{ref}. In this paper, we propose an improvement of SEQ by introducing a new way to build D_{ref} that is based on the highest suspect normal samples instead of abnormal samples. The new algorithm is applied to a number of benchmark datasets. The experimental results show that the proposed method provided better and more stable performance compared to the previous version of SEQ and six individual algorithms.

I. INTRODUCTION

In data mining, anomaly detection aims at identifying the observations which do not conform to an expected behavior. These observations are often referred to as anomalies or outliers. Anomaly detection methods are extensively used in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security. In recent years, ensemble methods for anomaly detection has received more attention in the research community [1], [2], [3], [4]. The objective of ensemble method is to combine various algorithms to create a robust algorithm for different problems. In anomaly detection, a particular algorithm may be well suited to the characteristics of one dataset and be effective in identifying anomalies of the specific application domain. However, that algorithm may not work well with other datasets whose properties are different. Ensemble methods, which use a variety of algorithms, may reduce the impact of such mismatch between an algorithm and an application, so as obtain better performance than the constituent algorithms alone and produce more robust results.

Regarding to component independence, ensemble algorithms are categorized into two classes: sequential ensemble and independent ensemble [5]. In independent ensembles, outlier detection are performed by using different algorithms or different portions of the data. The same algorithm may be used, with different initialization, parameter set or random seed. Then the results of those different applications are combined somehow to obtain a more robust detection effectiveness. In sequential ensembles, one or more outlier detection algorithms are sequentially applied on entire or some parts of data. Each application of the algorithms gives understanding to some aspects of the data, so using a modified algorithm or dataset would refine the detection results.

In a recent study, Zhao et al. proposed a new ensemble approach for anomaly detection based on sequential application of two algorithm (SEQ). The idea is to determine a reference dataset D_{ref} of the highest suspect abnormal samples based on the first algorithm and then the second algorithm is used to evaluate the final rank of data samples with respect to D_{ref}. The experimental results showed the promising performance on detection effectiveness of SEQ [6]. However, SEQ has a potential weakness in calculating the abnormal score based on the highest suspect abnormal subset D_{ref} especially when the second algorithm is the density based algorithm. Since D_{ref} is the suspect abnormal subset, the density of its samples may be very low. Subsequently, if the second algorithm is density-based, its evaluation for the density of all data samples based on D_{ref} will not be accurate (since evaluating the density of all samples with respect to D_{ref} (a low density subset) will result in a very small value).

In this paper, we proposed a variant of SEQ by selecting D_{ref} as the highest suspect of normal samples instead of abnormal samples. We evaluated this method on a larger range of public benchmark datasets with various values of main parameter k (the number of neighbor samples used to calculate the abnormal score for each data sample). The experimental results show that the proposed method achieved better performance than other tested algorithms.

The paper is organized as follows. The next section presents the background of the paper. The proposed sequential ensemble algorithm is presented in Section III. The tested datasets and the experimental settings are described in Section IV. The results are presented and discussed in Section V. Finally, Section VI concludes the paper and highlights some future work.

II. BACKGROUNDS

A. Sequential Application of two Algorithms

Zhao et al. [6] proposed a new ensemble method for anomaly detection that sequentially applies two substantially different algorithms (SEQ method). The main idea of this method is to correct the errors of the first algorithm by applying the second algorithm based on the result of the first algorithm. More detailedly, the process of SEQ includes the following two phase: In the first phase, an anomaly detection algorithm (the first algorithm) is applied on the whole dataset and provide the ranks for the abnormal degree of the samples, which are then used to select a subset D_{ref} that $|D_{ref}| = \alpha * 100\%|D|$ - suspected to contain all outliers. In the second phase, another algorithm (the second algorithm) then is applied on the whole D with reference to D_{ref} only. For example, if the second algorithm search for k nearest neighbors of an object, only its neighbors that in D_{ref} are taken into account. In other words, abnormal score of each sample o_i is calculated by applying the second algorithm on dataset $D_i = D_{ref} \cup \{o_i\}$ instead of dataset D. The pseudo-code of this algorithm (shorted as SEQ-A) is given in Algorithm 1 [6].

Algorithm 1 SEQ-A algorithm

Input: dataset D, detection algorithm A and B

Output: a list of rank associated with each of the objects in D

1: $ScoreList_A(D)$=Apply algorithm A on D.
2: $RankList_A(D)$=Assign each object in D a rank sorted by $ScoreList_A(D)$ in decreasing order.
3: D_{ref} = Identify the top $\alpha \times 100\%$ objects in $RankList_A(D)$
4: $ScoreList_{Final}(D)$ = For each object $o \in D$, apply algorithm B and get anomaly score calculated with reference to D_{ref}.
5: $RankList_{Final}(D)$ = Assign each object in D a rank sorted by $ScoreList_{Final}(D)$ in decreasing order.

As shown in Algorithm 1, the output of two algorithms are utilized by different ways. First algorithm's output is used to create $RankList_A$, which then is used to build a subset D_{ref} and does not directly contribute to the final output. Second algorithm's output is used to create the final output $RankList_{Final}$. Since the final output is rely not only on the second algorithm but also the subset D_{ref}, if D_{ref} is selected to well represent dataset's distinctive characteristics, the final output will be better.

SEQ-A was tested on 3 datasets including Packed Executables, KDD 99, Wisconsin and with several values of parameter k (2, 4, 10). The results shows that the sequential method achieved very good performance. Moreover, the authors also provided a procedure for selecting a best pair of algorithms for sequential method by using correlation coefficient between algorithms' output. According to Zhao et al., SEQ-A work best when the output of two selected algorithms are most different.

B. Individual Algorithms

In order to use ensemble algorithm, a set of single algorithms need to be selected first. In in this paper, we follow Zhao et al. in selecting the six following single algorithms.

LOF (Local Outlier Factor) [7] compares the local reachability density (lrd) of the test object, with those of the object's k nearest neighbors $N_k(p)$: a high value indicates outlierness.

$$LOF_k(p) = \frac{1}{|N_k(p)|} \sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}. \quad (1)$$

Local reachability density is defined as the inverse of average reachability distance from the neighbors of an object.

$$lrd_k(p) = 1 / \frac{\sum_{o \in N_k(p)} reach - dist_k(p \leftarrow o)}{|N_k(p)|}. \quad (2)$$

The reachability distance of object p with respect to object o is defined as $reach - dist(p \leftarrow o) = max\{k - dist(o), d(p, o)\}$ where $k - dist(o)$ denote the distance from o to its k_{th} nearest neighbor.

COF (Connectivity-based Outlier Factor) [8] compare the average chaining distance from an object to its k nearest neighbors $N_k(p)$ and the average of the average chaining distances from its k nearest neighbors to their own k nearest neighbors.

$$COF_k(p) = \frac{|N_k(p)| ac - dist_{N_k(p)}(p)}{\sum_{o \in N_k(p)} ac - dist_{N_k(o)}(o)}. \quad (3)$$

The concept of average chaining distance is based on two other concepts: Set Based Nearest Path (SBN-Path) and Set Based Trail (SBT). SNN-Path represents objects $p_i \in N_k(p)$ as a sequence $\{p_1, p_2, \ldots, p_r\}$ such that for all $1 \leq i \leq r - 1$, p_{i+1} is the nearest neighbor of set $\{p_1, \ldots, p_i\}$ in $\{p_{i+1}, \ldots, p_r\}$. A SBT is an ordered collection of $k - 1$ edges associated with a given SBT-Path such that the ith edge e_i connects the $(i + 1)^{th}$ point p_{i+1} in SBN-Path to one of the nearest earlier points in the path; each edge is assigned a weight proportional to the order in which it is added to SBT set. Average chaining distance of p is defined as the weighted sum of the lengths of the edges:

$$ac - dist_{N_k(p)}(p) = \frac{1}{r-1} \cdot \sum_{i=1}^{r-1} \frac{2(r-i)}{r} \cdot dist(e_i). \quad (4)$$

INFLO (INFLuenced Outlierness) [9] is a variant of LOF in which outlierness is measured based on both nearest neighbors and reverse nearest neighbor relationship. Jin et al. [9] defined Influential Space for object p as $IS_k(p) = N_k(p) \cup RN_k(p)$ where reverse nearest neighbor is defined as $RN_k(p) = \{q | q \in D, p \in N_k(q)\}$ and defined the density of an object as the inverse of its k-distance $den(p) = 1/k_{dist}(p)$. Influenced outlierness of an object p then is defined as:

$$INFLO_k(p) = \frac{\sum_{o \in IS_k(p)} den(o)}{|IS_k(p)|} / den(p). \quad (5)$$

RBDA (Ranking-based Detection Algorithm) [10] identifies outliers based on mutual closeness of an object and its

neighbors using ranks instead of distances. The outlierness of an object is defined as the average of its ranks among all nearest neighbors of its nearest neighbors

$$O_k(p) = \frac{\sum_{q \in N_k(p)} rank_q(p)}{|N_k(p)|}.$$ (6)

RADA (Rank with Averaged Distance Algorithm) [11] is a variant of RBDA which utilize the useful information contained in the distance of an object to its neighbors. The outlierness of object p is adjust by the average of the distances from the object to its neighbors.

$$W_k(p) = O_k(p) \cdot \frac{\sum_{q \in N_k(p)} d(p,q)}{|N_k(p)|}.$$ (7)

where $O_k(p)$ is defined as in Equation 6.

ODMR (Outlier Detection using Modified-Ranks) [11] is a variant of RBDA using objects' modified-ranks in order to overcome the effect of cluster density. In ODMR, all clusters are assigned weight 1, so each object $p \in C$ is assigned an equal weight = $1/|C|$. The modified-rank of an object p with respect to q is defined as the sum of weights of all objects within the circle of radius $d(q,p)$ centered at q. The outlierness of object p is measure by Equation 6 with modified rank of p replace for $rank_q(p)$.

III. METHODS

In this section, we proposed an extension of SEQ-A by introducing a new way to select D_{ref}. Instead of selecting D_{ref} based on the suspect abnormal samples, we select D_{ref} based on the suspect normal samples. We hypothesize that this method will be more robust compared to the SEQ-A especially when the second algorithm is the density based algorithm. The proposed algorithm is call sequential ensemble method based on normal dataset and shorted as SEQ-N. The pseudo-code of the SEQ-N is given in Algorithm 2.

Algorithm 2 SEQ-N algorithm

Input: dataset D, detection algorithm A and B

Output: a list of rank associated with each of the objects in D

1: $ScoreList_A(D)$=Apply algorithm A on D.
2: $RankList_A(D)$=Assign each object in D a rank sorted by $ScoreList_A(D)$ in increasing order.
3: D_{ref} = Identify the top $\alpha \times 100\%$ objects in $RankList_A(D)$
4: $ScoreList_{Final}(D)$ = For each object $o \in D$, apply algorithm B and get anomaly score calculated with reference to D_{ref}.
5: $RankList_{Final}(D)$ = Assign each object in D a rank sorted by $ScoreList_{Final}(D)$ in decreasing order.

Compared to Algorithm 1, there is only one line in Algorithm 2 that is different from Algorithm 1. In the second line of Algorithm 2 the samples in D is sorted in increasing order (from the highest suspect normal to the highest suspect

abnormal) instead of decreasing order in Algorithm 1. Algorithm SEQ-N will be compared with SEQ-A and six individual algorithms on a large number of popular datasets with a wide range of their parameters (k). The performance of the tested algorithms will be presented in the following section.

IV. EXPERIMENTAL SETTINGS

A. Selection of Pair of Algorithms for The Sequential Algorithm

The selection of individual algorithms has a profound impact to the performance of the ensemble methods. In this paper, we followed Zhao et al. [6] in selecting COF or LOF as the first algorithm and RADA for the second phase of sequential method. Subsequently, there are totally 4 sequential algorithms that were evaluated in this study, denote by SEQ-A.COF, SEQ-A.LOF, SEQ-N.COF, and SEQ-N.LOF.

The number of neighbors used for calculating abnormal score (k) in all individuals algorithm is varied from 1 to 40. After that, the best values for each algorithm was selected and the performance of the algorithms with the best value of k is reported as the final performance of the algorithm. The value of α in the ensemble algorithms was 20 for all ensemble versions. In other words, we selected $|D_{ref}| = 0.2|D|$.

B. Datasets

The lack of useful, publicly available benchmark data for outlier detection tasks has been been a challenge for evaluation the performance of algorithms. Recently, Campos et al. [12] collected, and made publicity a repository of anomaly detection benchmark datasets [1]. Since those datasets are publicly provided with full details of their preprocessing, they are chosen to evaluate the algorithms in this study. Moreover, Campos et al. also showed that, the performance of anomaly detection on the normalized datasets are often better than on unnormalized datasets. In this paper, nine normalized datasets that frequently used in the literature were selected for testing ensemble algorithms. The datasets and their characteristics were described in Table I. In nine datasets, there is only Lymphography that has categorical properties, the Lymphography_idf variant was chosen [2], in which categorical attribute is encoded as the inverse document frequency $IDF(t) = ln(|D|/f_t)$.

C. Evaluation Measures

To evaluate the performance of the tested algorithm, we use a well-know performance metric, ROC AUC. ROC AUC is based on a curve known as the Receiver Operating Characteristic (ROC). ROC AUC is a single value that summarized ROC curve, that defined as the area under the ROC curve (AUC). The ROC AUC value ranges between 0 and 1. A perfect ranking of the data objects would result in a ROC AUC value of 1, whereas an inverted perfect ranking would produce a value approaching 0. A random ranking of the data
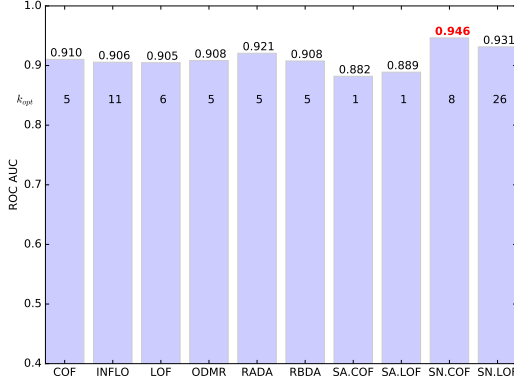

N^o	Dataset	Instance	Normal	Abnormal	Rate
1	Glass	214	205	9	4.2
2	Ionosphere	351	225	126	35.9
3	Lymphography	148	142	6	4.1
4	PenDigits	9868	9848	20	0.2
5	Shuttle	1013	1000	13	1.3
6	WBC	223	213	10	4.5
7	WDBC	367	357	10	2.7
8	WPBC	198	151	47	23.7
9	Waveform	3443	3343	100	2.9

Fig. 2. Result on dataset Waveform

Fig. 1. Result on dataset Ionosphere

Fig. 3. Result on dataset Lymphography

objects would result in a ROC AUC value close to 0.5. Since Campos et al. [12] confirmed the tendency that ROC AUC is expected to be less sensitive to variation in the number of true outliers than the other evaluation measures, we used ROC AUC as the main evaluation measure in this study.

V. RESULTS AND DISCUSSION

This section presents the performance of 10 tested anomaly detection algorithms on the datasets introduced in Sect. IV.

To compare methods according to their quality scores we firstly consider the best-case performance, selecting k for which the performance of a method on a data set is maximal (representing the optimistic case when the optimal value of k for a method is known in advance), then the case when the optimal values of k are not known.

Figure 1 to Figure 9 show the best-case performance these methods on nine datasets. In these figures, SA and SN are shorthanded for SEQ-A and SEQ-N ensemble methods. For example SN.COF is the sequential ensemble method using COF as the first algorithms to select D_{ref} of normal samples and then using RADA as the second algorithm with respect to D_{ref}. Moreover, the ROC AUC of the best algorithm on each dataset is presented by the red color. k_{opt} indicates the optimal values of parameter k at which each algorithm reaches its best performance.

It can be seen from Figure 1 to Figure 9 that SEQ-As (including SEQ-A.LOF and SEQ-A.COF) and SEQ-Ns (including SEQ-N.LOF and SEQ-N.COF) are often the best algorithms among all tested methods. Obviously, SEQ-As

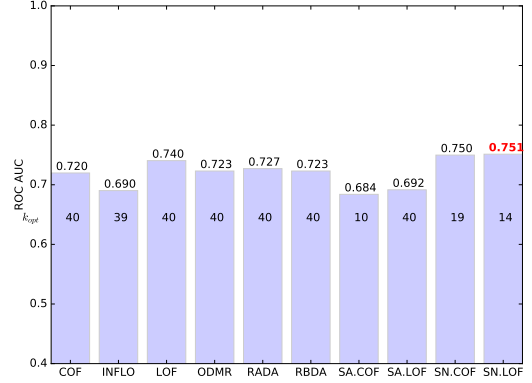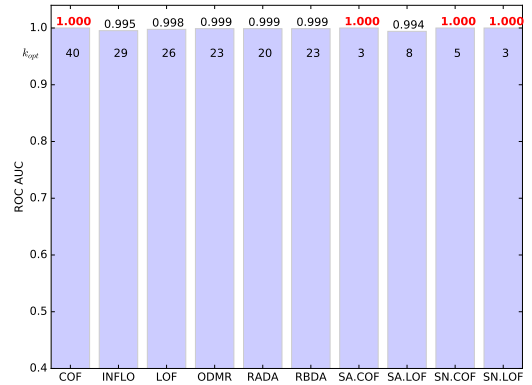achieved the best result on five datasets (Lymphography, WBC, Glass, WPBC and WDBC) while SEQ-Ns (SEQ-N.LOF or SEQ-N.COF) achieved the best result on four over nine tested datasets (Ionosphere, Waveform, Lymphography, WBC) and RADA achieved the best performance on only two datasets (Shuttle and PenDigits).

Comparing between two ensemble methods (SEQ-Ns and SEQ-As), we can see that the proposed ensemble methods (SEQ-Ns) are more robust than the previous version (SEQ-
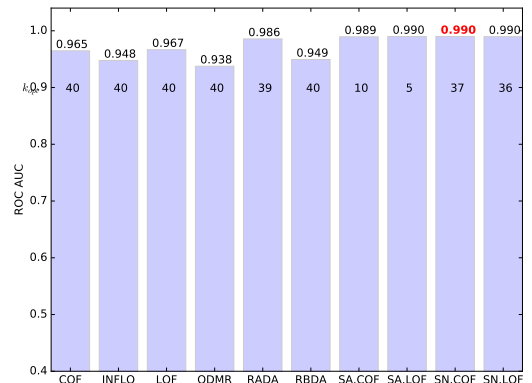

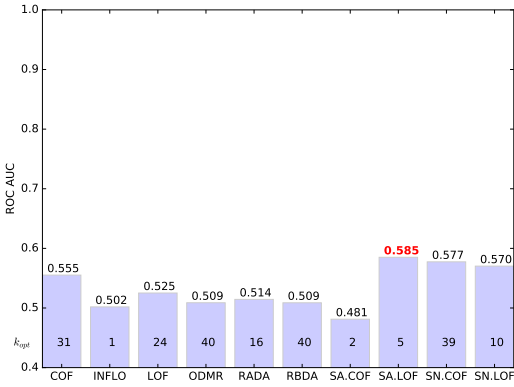Fig. 4. Result on dataset WBC

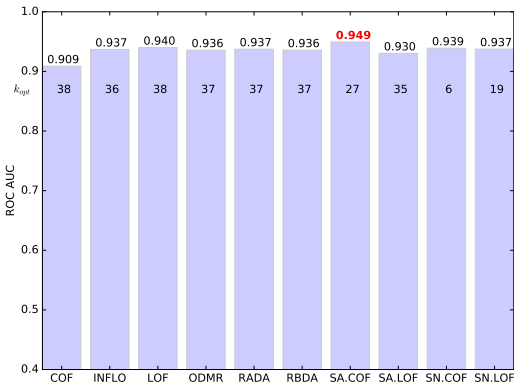Fig. 5. Result on dataset WPBC

Fig. 8. Result on dataset Shuttle
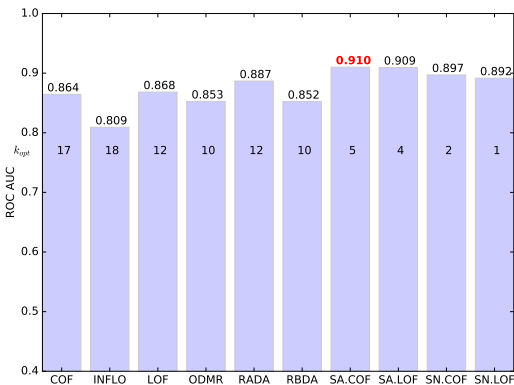

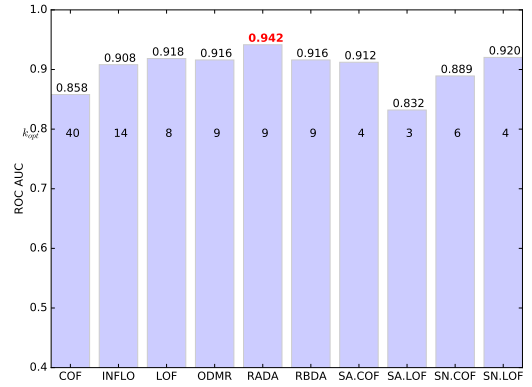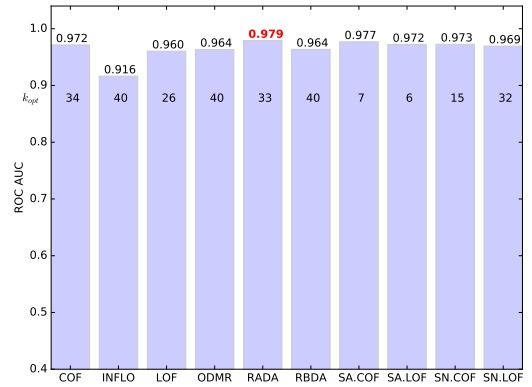Fig. 6. Result on dataset WDBC

Fig. 9. Result on dataset PenDigits

Fig. 7. Result on dataset Glass

As). This is presented by the fact that, on some problems, although, SEQ-Ns does not achieved the best result, their performance are still very convincing and their ROC AUC values are often very close to the best result on that dataset. Conversely, the performance of SEQ-As on some datasets is very unsatisfied (Ionosphere and Waveform). On these two datasets (Ionosphere and Waveform), SEQ-As obtained the worst results among all tested algorithms. Moreover, the performance of SEQ-Ns is also less sensitive regarding to the

selection of the algorithm in the first phase. Apparently, both versions of SEQ-N (SEQ-N.LOF and SEQ-N.COF) achieved good results while on some datasets, the performance of one of the versions of SEQ-A is very unconvincing (SEQ-A.COF on WPBC and SEQ-A.LOF on Shuttle).

Finally, comparing between ensemble algorithms with individuals algorithms, we can see that ensemble algorithms, especially SEQ-N, are often better all tested individuals. Moreover, the performance of SEQ-Ns is more robust than any other individual algorithms. On most dataset, SEQ-Ns are often among the three best algorithms.

Figure 10 presents the average value of ROC AUC over all datasets with respect to the choice of k (consider the case when the optimal values of k is not known). It can be observed that, two ensemble algorithms often achieved good performance with small values of k while the individuals algorithms only achieved good performance when k is large enough. There are two benefits of ensemble algorithms when they achieved good performance with relatively small value of k. First, it will be easier to tune this parameter for ensemble algorithms (since we only need to search for a smaller range of values). Second, this will help to reduce the computational time of ensemble algorithms since we will only need to calculate the score for each sample based on a smaller set. In the future, we will quantitatively investigate this.
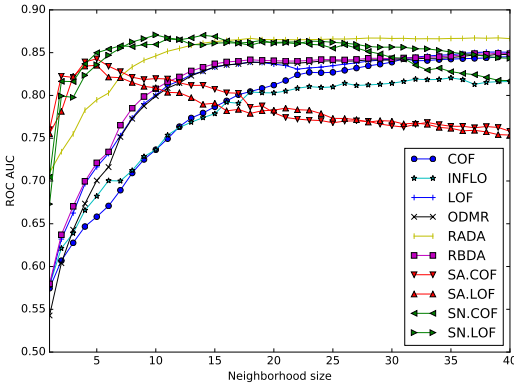
Fig. 10. Overall Result on nine datasets

Figure 10 also shows that SEQ-Ns are often much better than SEQ-As when considering the average results on all tested datasets. particularly, SEQ-Ns are much better than SEQ-As when k is greater than 3. Moreover, we can see that SEQ-Ns are the best algorithms among all tested algorithms and RADA is the only algorithm that achieved the performance close to SEQ-Ns.

Overall, the results in this section show that the proposed ensemble method (SEQ-N) helps to improve the performance of the previous version (SEQ-A) in detecting abnormal samples in the tested datasets. Moreover, SEQ-N also achieved the best result among all tested algorithms including six individual algorithms and four ensemble algorithms. Particularly, the performance of SEQ-N is more robust and less sensitive to the selection of the algorithm in the first phase than the previous algorithm SEQ-A.

VI. CONCLUSION

In this paper, we have proposed a novel sequential ensemble method (SEQ-N) for anomaly detection. The idea is using an anomaly detection algorithm result to select a subset of the dataset (D_{ref}) such that its members have the highest probability to be normal samples. After that, a second algorithm is used to calculate the final anomaly scores for all data samples with reference to only D_{ref}. The proposed algorithm was tested on nine widely used datasets and compared with the previous sequential ensemble algorithm by Zhao et al. [6] and six individual algorithms. The experimental results show that SEQ-N achieved the best results compared to all tested algorithms based on a reliable performance metric (ROC AUC).

There are some research areas for future work which arise from this paper. First, we would like to evaluate the effect of the sequential methods with different pair of base algorithms. In this paper, we followed Zhao et al. [6] by selecting LOF and COF for the first phase and RADA for the second phase of the algorithm. However, the combination of other algorithms including some recent reconstruction based methods [13], [14] may give better performance of SEQ-N. Second, it will be very interesting to study the effect of the size of D_{ref} to the

effectiveness of SEQ-N. Last by not least, we want to test the performance of SEQ-N on a wider range of datasets including the datasets recently recommended by Campos et al. [12].

REFERENCES

[1] H. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011*. SIAM / Omnipress, 2011, pp. 13–24.

[2] E. Schubert, R. Wojdanowski, A. Zimek, and H. Kriegel, "On evaluation of outlier rankings and outlier scores," in *Proceedings of the Twelfth SIAM International Conference on Data Mining*. SIAM / Omnipress, 2012, pp. 1047–1058.

[3] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander, "Subsampling for efficient and effective unsupervised outlier detection ensembles," in *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013*, 2013, pp. 428–436.

[4] S. Rayana and L. Akoglu, "Less is more: Building selective anomaly ensembles with application to event detection in temporal graphs," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, S. Venkatasubramanian and J. Ye, Eds. SIAM, 2015, pp. 622–630.

[5] C. C. Aggarwal, "Outlier ensembles: position paper," *SIGKDD Explorations*, vol. 14, no. 2, pp. 49–58, 2012.

[6] Z. Zhao, K. G. Mehrotra, and C. K. Mohan, "Ensemble algorithms for unsupervised anomaly detection," in *Current Approaches in Applied Artificial Intelligence - 28th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE*, 2015, pp. 514–525.

[7] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, W. Chen, J. F. Naughton, and P. A. Bernstein, Eds. ACM, 2000, pp. 93–104.

[8] J. Tang, Z. Chen, A. W. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD 2002*, ser. Lecture Notes in Computer Science, M. Cheng, P. S. Yu, and B. Liu, Eds., vol. 2336. Springer, 2002, pp. 535–548.

[9] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference, PAKDD 2006*, ser. Lecture Notes in Computer Science, W. K. Ng, M. Kitsuregawa, J. Li, and K. Chang, Eds., vol. 3918. Springer, 2006, pp. 577–593.

[10] H. Huang, K. Mehrotra, and C. K. Mohan, "Rank-based outlier detection," *Journal of Statistical Computation and Simulation*, vol. 83, no. 3, pp. 518–531, 2013.

[11] ——, "Algorithms for detecting outliers via clustering and ranks," in *Advanced Research in Applied Artificial Intelligence - 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012*, ser. Lecture Notes in Computer Science, H. Jiang, W. Ding, M. Ali, and X. Wu, Eds., vol. 7345. Springer, 2012, pp. 20–29.

[12] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Min. Knowl. Discov.*, vol. 30, no. 4, pp. 891–927, 2016.

[13] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 2014, p. 4.

[14] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," SNU Data Mining Center, Tech. Rep., 2015.