ICIC Express Letters
Volume **14**, Number **5**, May **2020**

 ISSN 1881-803X
pp. 515–520

A STUDY ON EVALUATION MEASURES FOR UNSUPERVISED OUTLIER DETECTION

Sunmin La[1] and Nam-Wook Cho[2,*]

[1]Department of Data Science
Graduate School
[2]Department of Industrial and Information Systems Engineering
Seoul National University of Science and Technology
232, Gongneung-ro, Nowon-gu, Seoul 01811, Korea
*Corresponding author: nwcho@seoultech.ac.kr

Received November 2019; accepted February 2020

Abstract. *Outlier detection is a data analysis method based on data mining techniques and is used to identify outlying observations which might have significance in a dataset. Research on outlier detection, however, has mainly focused on supervised approaches, which require labeled training and test datasets. Unsupervised approaches are more appropriate for many applications such as network intrusion detection and fraud detection, but the suitability of these methods to determine the degree of outlierness of a dataset has not been fully addressed because the ground truth is usually unavailable. In this paper, evaluation measures for unsupervised outlier detection, which can effectively measure the outlierness of a dataset, are proposed. To verify the effectiveness of the proposed methods, experiments were conducted with University of California Irvine machine learning datasets using a k-nearest neighbors (k-NN) algorithm.*

Keywords: Unsupervised outlier detection, k-nearest neighbors (k-NN), Gini index, External measure, Outlierness

1. **Introduction.** Outlier detection is an important task for various data mining applications [1]. It aims at finding abnormal observations that can be considered to be inconsistent with respect to the remainder of a dataset [2-4]. An outlier can be defined as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [4].

Outlier detection has been playing important roles in various applications including intrusion detection, fraud detection and health care [6-8]. The application of outlier detection has been further extended to trajectory and moving object detection [9], emerging topic detection [10], and temporal data detection [1].

Research into outlier detection, however, has mainly focused on supervised approaches, which require labeled training and test datasets [11]. The computational efficiency and accuracy of supervised outlier detection algorithms have been of much interest to researchers [12]. In contrast, unsupervised outlier detection means that labeled data is unavailable and, consequently, no ground truth is available for the assessment of the quality of an algorithm [13]. Many applications, such as network intrusion detection, should process unlabeled data in an unsupervised environment. Consequently, the evaluation of the models often depends on the subjective judgement of researchers [13] because there is no ground truth for the measurement of precision, accuracy, or recall.

Thus, the need for external measures to determine the outlierness of a dataset has been highlighted but not been fully addressed. In this paper, a Gini-index-based evaluation measurement for unsupervised outlier detection, which can effectively measure the degree

DOI: 10.24507/icicel.14.05.515

of outlierness of a dataset, is proposed. To verify the effectiveness of the proposed methods, experiments were conducted with ten University of California Irvine (UCI) machine learning datasets using a k-nearest neighbor algorithm, one of the most popular outlier detection algorithms [14].

Typical outputs of an outlier detection algorithm are 1) a label indicating that an instance is an outlier or not and 2) a score indicating the degree of abnormality of a dataset [11,15]. The main focus of our research is to score the degree of abnormality or outlierness of a dataset in an unsupervised environment. The questions addressed in this paper are as follows: 1) for a given dataset, do outliers exist and; 2) if outliers exist in a dataset, what is the degree of their outlierness?

The rest of this paper is organized as follows. Section 2 provides a research framework, along with methodologies and datasets used in the paper. Section 3 explains the results of the experiments conducted with the UCI datasets. Finally, Section 4 discusses the benefits and limitations of our research.

2. **Methods.** In this paper, Gini-index-based evaluation methods for unsupervised outlier detection are suggested and tested using a k-nearest neighbors (k-NN) algorithm. The data used in this research were obtained from the UCI Machine Learning Repository [16]. Table 1 summarizes the datasets used in the paper. Among the UCI datasets used for classification, ten commonly used datasets in outlier detection were selected for the experiments. The subsequent sections explain the details of the methods used in this paper.

Table 1. UCI datasets

No.	Dataset	Normal class	Outlier class	# of instances	# of features	Outlier percent (%)
1	PageBlock	1	2, 3, 4, 5	5,473	10	560 (10.2%)
2	Cardio	1 (normal)	3 (pathologic)	1,831	21	176 (9.6%)
3	HTRU2	0	1	17,898	8	1,639 (9.2%)
4	Shuttle	1	2, 3, 4, 5, 6, 7	12,345	9	867 (7%)
5	Wilt	N (others)	W	4,839	5	261 (5.4%)
6	Glass	Others	6	214	7	9 (4.2%)
7	Waveform	Others	0*	3,443	21	100 (2.9%)
8	WDBC	Benign	Malignant*	367	30	10 (2.7%)
9	Annthyroid	3, 2	1	3,772	21	93 (2.5%)
10	PenDigits	Others	4*	9,868	16	20 (0.2%)

*Downsampled for experiment.

2.1. **k-nearest neighbor.** The k-NN approach is based on the distance between data points. It is one of the most commonly used methods for outlier detection and is often preferred in practical applications [11]. The k-NN approach assumes that normal data points have close neighbors, whereas outlying data points are located relatively far from their neighbors [8,14].

The k-NN unsupervised outlier detection algorithm is a straightforward way of detecting outliers. A data point is identified as an outlier if it is located far from its neighbors. Among the methods for calculating the distance, the Euclidean distance, Mahalanobis distance, and Minkowski distance are commonly used [14]. In this study the Minkowski distance was used because it is considered as a generalization of the Euclidean and Manhattan distances. The Minkowski distance of order p between two data points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ is defined as

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}. \tag{1}$$

2.2. **Gini index.** Entropy has been used to measure the impurity of a dataset [3], but the Gini index considers impurity and inequality as well. The Gini index is commonly used in economics when measuring inequality of income or wealth. It is also used in decision trees when splitting branches. If a dataset S contains m classes, the Gini index is defined as follows [17]:

$$Gini(S) = 1 - \sum_{j=1}^{m} P_j^2, \tag{2}$$

where P_j is the relative frequency of class j in S. Note that the k-NN score was rounded to the second decimal point when calculating the Gini index.

2.3. **Procedures.**

1) Preprocessing of datasets

Before the experiments, data points with missing values were removed. The values of the numerical variables were standardized so that distance measures were not overly influenced by certain variables.

2) Application of k-NN algorithm and calculation of Gini index

For each dataset, k-NN scores of all data points are obtained. Then, a Gini index is calculated for each dataset. A high Gini index means that the k-NN scores of a dataset are not evenly distributed, indicating high degree of outlierness of the dataset.

3) Cut-off point

For each dataset, all data points are sorted in descending order of k-NN scores. Let $Z = \{z_1, z_2, \ldots, z_n\}$ be an ordered list of k-NN scores of data points $(i = 1, 2, \ldots, n)$ in dataset S. Then, a cut-off point is defined as follows: Let $d_i = z_i - z_{i+1}$; then cut-off point c of S is the smallest i such that $d_i < \delta$, where δ is a small positive real number that determines the cut-off point.

A cut-off point can be used to identify potential outliers; data points whose k-NN score is greater than the k-NN score of a cut-off point can be considered as outliers because their k-NN scores are significantly higher than remainder of the dataset. A cut-off point can also be used to evaluate the outlierness of a dataset. If a cut-off point is close to zero, it implies that a small number of outliers may exist in the dataset. Figure 1 illustrates the cut-off point.

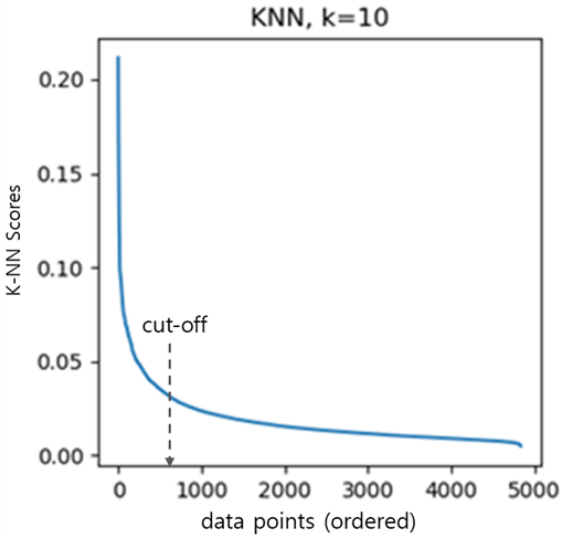

FIGURE 1. Cut-off point

4) Calculation of outlierness evaluation metrics

In this paper, three evaluation metrics are proposed. The Gini index and cut-off point are utilized to evaluate the outlierness of a dataset.

a. Outlierness1 (%)

For a given dataset, outlierness1 is defined as the percentage of the number of potential outliers to the total number of data points in a dataset. Note that potential outliers are identified by using a cut-off point.

b. Outlierness2 (%)

For a given dataset, outlierness2 is defined as the percentage reduction of the Gini index after removing potential outliers from the dataset.

c. Outlierness multipliers

The outlierness multiplier is defined by a ratio (outlierness2/outlierness1) that represents the amount of outlierness contributed by the potential outliers.

3. **Results and Discussion.** Table 2 summarizes the experiments conducted with the ten UCI datasets in Table 1. Some datasets (Annthyroid and PenDigits) have a relatively small number of potential outliers compared with the other datasets. For each dataset, the Gini indices were compared after removing the potential outliers. The percent decrease in the Gini index (outlierness2) suggests that the impurity or inequality of the datasets has been improved significantly after removing the outliers. Concerning the outlierness multiplier, the PenDigits dataset exhibits the highest value among those of all datasets, indicating that its outliers account for a high degree of outlierness.

Table 2. Experiments results

No.	Dataset	Outlierness1 (cut-off percentage)	Gini index	Gini index (after removing potential outliers)	Outlierness2 (percent reduction in Gini index)	Outlierness multiplier
1	PageBlock	12.8%	0.0680	0.0461	32.2%	2.52
2	Cardio	13.7%	0.0605	0.0346	42.8%	3.12
3	HTRU2	10.0%	0.0417	0.0272	34.8%	3.48
4	Shuttle	8.1%	0.0771	0.0495	35.8%	4.42
5	Wilt	8.0%	0.0537	0.036	33.0%	4.13
6	Glass	11.7%	0.1628	0.0895	45.0%	3.85
7	Waveform	7.3%	0.0329	0.0258	21.6%	2.96
8	WDBC	13.6%	0.0684	0.0471	31.1%	2.29
9	Annthyroid	5.3%	0.2613	0.1575	39.7%	7.49
10	PenDigits	2.5%	0.0539	0.0429	20.4%	8.16

The outlierness metrics can be used to analyze the dataset's characteristics further, as shown in Figure 2 and Table 3. In this paper, an outlierness quadrant is suggested to illustrate the degree of outlierness of a dataset. As shown, the majority of datasets are located in the fourth quadrant, which means that there are large number of outliers that contribute to a relatively small degree of outlierness. This contrasts with the datasets in the second quadrant, Annthyroid and PenDigits, where a small number of outliers contribute to a relatively high degree of outlierness. Consequently, the datasets in the second quadrant have a relatively high degree of outlierness.

4. **Conclusion.** In this paper, a method of evaluation of unsupervised outlier detection has been proposed. Experiments were conducted with ten UCI datasets. The results show that the proposed metrics effectively measure the outlierness of datasets. It has been shown that the proposed measures can overcome the subjective nature of existing

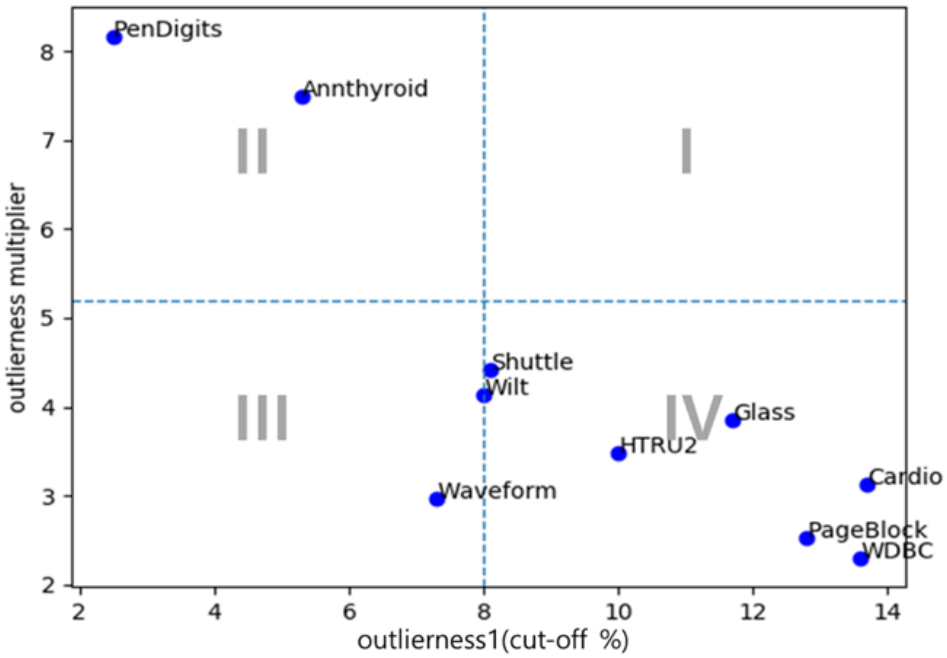

FIGURE 2. Outlierness quadrant

TABLE 3. Outlierness quadrant

Quadrant	Description	Datasets
First	A dataset has a large number of outliers whose degree of outlierness is high.	–
Second	A dataset has a small number of outliers whose degree of outlierness is high.	Annthyroid, PenDigits
Third	A dataset has a small number of outliers whose degree of outlierness is small.	Waveform
Fourth	A dataset has a large number of outliers whose degree of outlierness is small.	Shuttle, HTRU2, Glass, Cardio, PageBlock, WDBC, Wilt

evaluation measures, which is the main contribution of our paper. This paper also demonstrates the usefulness of the Gini index in evaluating the outlierness of datasets. Although the Gini index is known to be an effective measure for both impurity and inequality, it has received little attention from outlier detection models. This study showed that the Gini index could be an appropriate metric for evaluating the outlierness of a dataset, which is another contribution of our study.

Despite the contribution of our study, it has some limitations. Although the experimental results are consistent, only ten datasets were examined; thus, the study set might be insufficient for generalization. Consequently, more extensive experiments using real datasets in an unsupervised environment should be conducted to reinforce our findings.

Although our evaluation method is less subjective than existing methods, it is not free from subjective judgement. The determination of a threshold value (δ) was still required when choosing a cut-off point. Thus, the identification of a more rigorous approach for determining a cut-off point would be a suitable future research topic.

REFERENCES

[1] S. E. Benkabou, K. Benabdeslem and B. Canitia, Unsupervised outlier detection for time series by entropy and dynamic time warping, *Knowledge and Information Systems*, vol.54, no.2, pp.463-486, 2018.

[2] S. Kim, N. W. Cho, B. Kang and S.-H. Kang, Fast outlier detection for very large log data, *Expert Systems with Applications*, vol.38, no.8, pp.9587-9596, 2011.

[3] A. Koufakou, E. G. Ortiz, M. Georgiopoulos, G. C. Anagnostopoulos and K. M. Reynolds, A scalable and efficient outlier detection strategy for categorical data, *The 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp.210-217, 2007.

[4] Z. He, S. Deng and X. Xu, An optimization model for outlier detection in categorical data, in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang and G.-B. Huang (eds.), Springer Berlin Heidelberg, 2005.

[5] D. M. Hawkins, *Identification of Outliers*, Springer, 1980.

[6] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang and X. He, Generative adversarial active learning for unsupervised outlier detection, *IEEE Trans. Knowledge and Data Engineering*, 2019.

[7] P. Kang, K. Kim and N. W. Cho, Toll fraud detection of voip services via an ensemble of novelty detection algorithms, *International Journal of Industrial Engineering*, vol.22, no.2, 2015.

[8] V. Hautamaki, I. Karkkainen and P. Franti, Outlier detection using k-nearest neighbour graph, *Proc. of the 17th International Conference on Pattern Recognition (ICPR)*, pp.430-433, 2004.

[9] J. Mao, P. Sun, C. Jin and A. Zhou, Outlier detection over distributed trajectory streams, *Proc. of the 2018 SIAM International Conference on Data Mining*, pp.64-72, 2018.

[10] H. Soleimani and D. J. Miller, ATD: Anomalous topic discovery in high dimensional discrete data, *IEEE Trans. Knowledge and Data Engineering*, vol.28, no.9, pp.2267-2280, 2016.

[11] M. Goldstein and S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PloS One*, vol.11, no.4, 2016.

[12] G. O. Campos, A. Zimek, J. Sander, R. J. 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 Mining and Knowledge Discovery*, vol.30, no.4, pp.891-927, 2016.

[13] C. C. Aggarwal, Outlier ensembles: Position paper, *ACM SIGKDD Explorations Newsletter*, vol.14, no.2, pp.49-58, 2013.

[14] M. A. Pimentel, D. A. Clifton, L. Clifton and L. Tarassenko, A review of novelty detection, *Signal Processing*, vol.99, pp.215-249, 2014.

[15] A. Zimek, R. J. Campello and J. Sander, Ensembles for unsupervised outlier detection: Challenges and research questions. A position paper, *ACM SIGKDD Explorations Newsletter*, vol.15, no.1, pp.11-22, 2014.

[16] D. Dua and C. Graff, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, CA, http://archive.ics.uci.edu/ml, 2019.

[17] W. Du and Z. Zhan, Building decision tree classifier on private data, *Proc. of the IEEE International Conference on Privacy, Security and Data Mining*, pp.1-8, 2002.