Journal of Computer Science and Cybernetics, V.32, N.3 (2016), 259–272
DOI 10.15625/1813-9663/32/3/8455

AN EVALUATION METHOD FOR UNSUPERVISED ANOMALY DETECTION ALGORITHMS

HUY VAN NGUYEN[1], TRUNG THANH NGUYEN[2], AND QUANG UY NGUYEN[2]

[1]*Institute of Information Technology, Vietnam Academy of Military Science and Technology;*
[2]*Faculty of IT, Le Quy Don Technical University;*
[1]*vannguyenhuy.vn@gmail.com;* [2]*quanguyhn@gmail.com,* [2]*trungthanhnt@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. These techniques have been successfully applied to many real world applications such as fraud detection for credit cards and intrusion detection in network security. However, there are very little research relating to the method for evaluating the goodness of unsupervised anomaly detection techniques. In this paper, the authors introduce a method for evaluating the performance of unsupervised anomaly detection techniques. The method is based on the application of internal validation metrics in clustering algorithms to anomaly detection. The experiments were conducted on a number of benchmarking datasets. The results are compared with the result of a recent proposed approach that shows that some proposed metrics are very consistent when being used to evaluate the performance of unsupervised anomaly detection algorithms.

Keywords. Anomaly detection, evaluation, clustering validation.

1. INTRODUCTION

Detecting anomaly has received great attention from the research community in machine learning [6]. Anomaly detection aims at finding samples in data that do not follow the expected behavior. These samples are often referred to as anomalies or outliers. These two terms are often used interchangeably. Anomaly detection techniques have extensively been applied to a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security [18].

Since the first study of anomaly detection by statisticians in the early 19th century, there has been a variety of anomaly detection techniques developed for diverse application domains [6]. Regarding the availability of labeled data, anomaly detection techniques are classified into three classes: Supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection. Among them, unsupervised anomaly detection are the most popular techniques and they have been applied to a wide range of problems [6].

In unsupervised anomaly detection, since labeled data is not available, evaluating the accuracy of detection methods has been a constant challenge in data mining research [31]. So far, the performance of unsupervised anomaly detection techniques has often been tested by using labeled data sets. In other words, the labels are not used by the algorithms during the training process, but only for evaluating their results [5]. This method is often referred to as external evaluation approach.

The downside of the external methods is that they are not applicable to many real world problems where the labeled data is not available. To the best of our knowledge, there has only been two published research on the approach for evaluating the accuracy of unsupervised anomaly detection techniques by Marquest et al. and Nguyen et al. [15, 17]. They proposed the idea of using classification algorithm to measure the performance of unsupervised anomaly detection techniques. Their method assumes that abnormal samples are often farther from normal samples and can therefore be more easily separated from other samples. Based on this assumption, they applied a classification algorithm on the output of anomaly detection and used the accuracy of the classification algorithm as the indicator for the performance of detection techniques.

The drawback of the approach in [15] and [17] is that it requires to execute one more algorithm (logistic regression in [17] and kernel based classification in [15]) on the top of anomaly detection techniques. Subsequently, this method may be considerably computational expensive. Moreover, the results may also depend on the selected classification algorithm and the parameters settings for the classification algorithms. Various classification algorithms with different settings may lead to significantly different results of the performance of anomaly detection methods.

In this paper, a new method for evaluating the accuracy of unsupervised anomaly detection approaches is introduced. The main contributions of the paper are:

- The use of internal validation metrics in clustering for measuring the performance of unsupervised anomaly detection algorithms proposed.

- Ten validation metrics are tested on some benchmarking abnormal datasets and compared with the method that used logistic regression [17].

- The experimental results show that one of the tested metrics is better than logistic regression when used for measuring the accuracy of unsupervised anomaly detection algorithms.

The paper is organized as follows. The next section introduces some popular anomaly detection techniques. The method for measuring the performance of unsupervised anomaly detection algorithms is discussed in Section 3. Section 4 presents ten internal metrics that will be examined in this paper. Section 5 describes the tested datasets and the experimental setup. The results of testing evaluation metrics are presented in Section 6. Section 7 concludes the paper and highlights some future work.

2. BACKGROUNDS

Anomaly detection has been the topic of a large number of research. For a comprehensive review of the research on anomaly detection, the readers are recommended to read [6]. In this section, some well-known anomaly detection approaches are described. Based on the extent to which the labeled data is available, anomaly detection techniques are categorized into the following three classes [6].

Supervised anomaly detection: These methods rely on the assumption that labeled instances for both normal and anomaly class are available during the training process. The objective is to learn a predictive model for normal versus abnormal classes. The resulting model is then used to determine which class an unseen data sample belongs to [23, 27]. Although, the accuracy of supervised methods is often higher than other approaches, labeling data is often strenuous and expensive. Subsequently, supervised anomaly detection techniques have not been applied as frequently as unsupervised methods.

Semi-Supervised anomaly detection: Semi-supervised techniques assume that there is only one class of instances (often normal class) in the training data. The typical approach is to construct a model for normal behavior, and use the model to identity anomalies in the test data. In the testing phase, if an unseen sample is not recognized by the learnt model, then this sample is considered as anomaly. Popular anomaly detection techniques based on one class learning include one-class Kernel Fisher Discriminants [20] and one class support vector machine [19]. These methods are more widely applicable than supervised techniques since abnormal instances are not required in the training phase.

Unsupervised anomaly detection: The techniques that operate in unsupervised mode do not require labeled samples for both classes. Thus they are the most widely applicable techniques. The techniques in this category assume that normal instances are far more frequent than anomalies in the dataset and that abnormal samples are often significantly different from the normal samples. To date, a large number of unsupervised anomaly detection approaches have been developed [6]. Among them, nearest neighbor based techniques, clustering based techniques and statistical techniques have been widely applied.

Nearest neighbor based techniques assume that the density in normal region is higher than in abnormal region. In other words, the abnormal samples are supposed to be isolated or lied in the region of sparse density. Therefore, the distance of a data instance to its k^{th} nearest neighbor can be considered as the anomaly score. If this distance is too high then the data sample is suspected to be anomaly [3, 30]. The advantage of nearest neighbor based techniques is that they are unsupervised in nature. Nevertheless, the computational complexity of the techniques in testing phase is often high ($O(N^2)$, N is the number of samples). This hinders the application of nearest neighbor based techniques to some real world applications where time constraint is important.

Clustering based techniques make the assumption that normal data instances belong to large clusters, while anomalies belong to small clusters. The techniques use clustering algorithms to divide the dataset into a number of clusters and report any instance that does not belong to any cluster or belongs to the clusters with a small number of samples as anomalous [14, 29]. Similar to nearest neighbor based techniques, clustering based techniques can operate fully in unsupervised mode. However, the computational complexity for clustering the data is also challenging in testing phase especially for the algorithms such as hierarchical clustering [24] where the complexity is $O(N^2)$.

Statistical anomaly detection techniques rely on the assumption that normal data instances are generated by a probabilistic model. The training process aims to learn the parameters of the probabilistic model. In the testing phase, the methods declare any sample with low probability of being generated from the learnt model as anomalous [2, 9]. The advantage of statistical approaches is that the complexity of fitting data is low (often linear). Consequently, statistical approaches have extensively been used in variety of real world problems particularly when the data volume is high. However, statistical approaches are based on the assumption that the data is generated from a particular distribution. If this assumption does not hold, the results of statistical methods may not be robust.

3. METHODS

The evaluation method introduced in this paper relies on the assumption that if an anomaly detection technique performs well on a dataset then the technique will separate the normal and abnormal samples into different clusters. In other words, let C_1 and C_2 be the normal and abnormal sets/clusters determined by applying an anomaly detection technique A on the data D, then we can see that: the performance of A on D is good if C_1 and C_2 are well separated and the performance of A on D is not good if C_1 and C_2 are not well separated (i.e. some normal samples are mixed in the abnormal clusters and vice versa). Therefore, it will be relevant to use the metrics for evaluating the performance of clustering algorithms (i.e. to evaluate if C_1 and C_2 are well separated) to measure the quality of anomaly detection techniques. The objective of this paper is to examine whether the validation metrics used in clustering could be applied for measuring the performance of anomaly detection approaches.

In data mining, clustering validation has long been recognized as one of the vital issues to the success of clustering applications [16]. Clustering validation is based on the metric to evaluate the goodness of clustering results. Clustering validation metrics can be categorized into two main types: external clustering validation and internal clustering validation. External validation metrics use external information (for example the data labels) that is not presented in clustering process to evaluate the extent to which the clustering structure discovered by an algorithm matches to the external structure. On the other hand, internal measures evaluate the goodness of a clustering structure without using any external information.

In this paper, we focus only on examining the internal validation metrics since they are more suitable for the unsupervised anomaly detection algorithms. In order to examine these metrics, a number of benchmarking datasets were selected. The labels of the chosen datasets are available but they are not used when calculating the tested validation metrics. This is to mimic the scenario in many real-world applications such as bank transfer, online game, etc. where the data label is not available. The data label is used only for swapping some data samples between normal and abnormal cluster to simulate the situation in which an anomaly detection technique inaccurately identifies some normal samples as abnormal samples and vice versa.

For each dataset, the normal and abnormal samples are divided into two clusters (the normal and the abnormal cluster). The validation metrics are then measured on these two clusters and the obtained value is referred to as t_0. Next, a number of abnormal samples is selected and swapped to the normal cluster. Correspondingly, the same number of normal samples are swapped to the abnormal cluster. The validation metrics are again applied to two new formed clusters (the normal and abnormal clusters that were formed after swapping some data samples). The swapping process is repeated 10 times with the number of swapped samples is varied from 10% to 100% of abnormal samples. The value of the validation metrics calculated on the new formed clusters is referred to as $t_1, ..., t_{10}$, respectively.

Assuming that, for a specific validation metric, the greater value presents the better clustering result. Then we expect its value will gradually decrease from t_0 to t_{10}. Conversely, if the smaller value of the metric indicates the better clustering result, then we expect that the value from t_0 to t_{10} is gradually increased. To measure the correlation between the validation metric result and the percent value of the swapped samples in the dataset, we calculate the Pearson correlation coefficient of two vectors: The first vector is $T = \{t_0, t_1, ..., t_{10}\}$ and the second vector is $K = \{0, 1, ..., 10\}$.

In statistics, the correlation between sets of data is a measure of how well they are related [1]. The most common measure of correlation in statistics is the Pearson correlation. This measure shows the linear relationship between two sets of data. If the value of Pearson correlation coefficient is close to 1 or -1 then two sets of data are highly correlated. In this case, the tested validation metric is good for measuring the performance of anomaly detection techniques. However, if the correlation coefficient is close to zero, then two datasets weakly correlated and the metric is not reliable for measuring the quality of detection approaches. In the following section, we will present 10 internal clustering validation metrics that are applied to measuring the goodness of anomaly detection techniques.

4. CLUSTERING VALIDATION METRICS

This section presents ten clustering validation metrics that will be tested for the evaluation of the goodness of anomaly detection. Since the target of clustering is ensuring objects within each cluster similar and objects in different clusters distinct [25], most of internal clustering validation measures are based on two criteria that are compactness and separation. Compactness measures the level of differently or closely related between the samples in the same cluster. Separation measures how a cluster is distinct or well-separated from other ones. The rest of this section will briefly present ten validation metrics. For the sake of the presentation, some important notations are shown in Table 1. In six out of ten tested metrics including RS, H, CH, S, I, D, the greater values present the better clustering result. Conversely, for four metrics (STD, DB, XB, SD), the smaller values mean the obtained clusters are better separated.

- The **root-mean-square standard deviation (*STD*)** [22] is the square root of the pooled sample variance of all the attributes. STD takes homogeneous level of the formed clusters into account[10] by summing up them and then normalizing the result.

$$RMSSTD = \left(\frac{\sum_{i=1}^{NC} \sum_{x \in C_i} \|x - c_i\|^2}{P \cdot \sum_{i=1}^{NC} (n_i - 1)} \right)^{\frac{1}{2}}. \tag{1}$$

- The **R-squared (*RS*)** [22] is the ratio of sum of squares between clusters to the total sum of squares of the whole data set. RS measures the degree of homogeneity between clusters [10].

$$RS = \frac{\sum_{x \in D} \|x - c\|^2 - \sum_{i=1}^{NC} \sum_{x \in C_i} \|x - c_i\|^2}{\sum_{x \in D} \|x - c\|^2}. \tag{2}$$

Table 1. Notation in clustering validation measure

Notation	Meaning
D	Data set
n	Number of objects in data set
c	Center of D
P	Number of attributes in data set
p	Contrast factor (taken $p = 2$)
NC	Number of clusters
C_i	The i_{th} Cluster
n_i	Number of objects in C_i
c_i	Center of C_i
$\sigma\left(C_i\right)$	Variance vector of C_i
$d\left(x,y\right)$	The distance between x and y
$\left\|X_i\right\|$	$\left(X_i^T \cdot X_i\right)^{\frac{1}{2}}$

- The **Modified Hubert statistic (*H*)** [13] evaluates the clustering quality the correlation between two square matrices of the same size. The first one is the proximity matrix of all objects in the data set and the second is the proximity matrix of the clusters' centers to which each object belongs.

$$H = \frac{2}{n\left(n-1\right)} \cdot \sum_{x\in D}\sum_{y\in D} d\left(x,y\right) d_{x\in C_i, y\in C_j}\left(c_i, c_j\right). \tag{3}$$

- The **CalinskiHarabasz index (*CH*)** evaluates the cluster solution based on the average between- and within-cluster sum of squares [4].

$$CH = \left[\sum\nolimits_{i=1}^{NC} n_i d^2\left(c_i, c\right) / \left(NC-1\right)\right] \Big/ \left[\sum\nolimits_{i=1}^{NC}\sum\nolimits_{x\in C_i} d^2\left(x, c_i\right) / \left(n - NC\right)\right]. \tag{4}$$

- The **Index *I* (*I*)** [16] takes the maximum distance between cluster centers as separation, and the sum of distances between objects and their cluster center as compactness.

$$I = \left(\frac{1}{NC} \cdot \frac{\sum_{x\in D} d\left(x,c\right)}{\sum_{i=1}^{NC}\sum_{x\in C_i} d\left(x, c_i\right)} \cdot \max\nolimits_{i,j} d\left(c_i, c_j\right)\right)^p. \tag{5}$$

- The **Dunns index (*D*)** [8] uses the minimum pair-wise distance between objects in different clusters as the inter-cluster separation and the maximum cluster diameter as the intracluster compactness. The index's value is ratio of the inter-cluster separation to the intra-cluster compactness.

$$D = \frac{\min_{1\le i\le NC}\min_{1\le j\le NC, j\ne i}\min_{x\in C_i, y\in C_j} d\left(x,y\right)}{\max_{1\le k\le NC}\max_{x\in C_k, y\in C_k} d\left(x,y\right)}. \tag{6}$$

- The **Silhouette index ($\boldsymbol{S}$)** [21] measures clustering partition based on the dissimilarity of each instance to its cluster's instances, and to its 'neighbor' cluster's instances.

$$S = \frac{1}{NC}\sum_{i=1}^{NC}\left[\frac{1}{n_i}\sum\nolimits_{x\in C_i}\frac{b(x)-a(x)}{\max\left[b(x), a(x)\right]}\right], \tag{7}$$

$$a(x) = \frac{1}{n_i-1}\sum\nolimits_{y\in C_i, y\neq x} d(x,y),$$

$$b(x) = \min_{j,j\neq i}\left[\frac{1}{n_j}\sum\nolimits_{y\in C_j} d(x,y)\right].$$

- The **DaviesBouldin index ($\boldsymbol{DB}$)** [7] is average of cluster's similarities. The similarity of each cluster is defined as the maximum value of its similarities to other clusters.

$$DB = \frac{1}{NC}\sum_i \max_{j,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)\right]\Big/ d(c_i,c_j)\right\}. \tag{8}$$

- The **Xie-Beni index ($\boldsymbol{XB}$)** uses the minimum square distance between cluster centers as intercluster separation and the mean square distance between each data object and its cluster center as the intracluster compactness. This index is defined as the ratio of the compactness to the separation [28].

$$XB = \frac{\sum_{i=1}^{NC}\sum_{x\in C_i} d^2(x,c_i)\Big/ n}{\min_{i,j} d^2(c_i,c_j)}. \tag{9}$$

- The last metric is ***SD*** **index ($\boldsymbol{SD}$)** [11] the summation of two terms: the average scattering and the total separation of clusters. The former evaluates compactness based on variances of clusters and dataset, and the latter evaluates separation difference based on distances between cluster centers.

$$SD = Dis(NC_{max})\cdot Scat(NC) + Dis(NC), \tag{10}$$

$$Scat(NC) = \frac{\sum_{i=1}^{NC}\left\|\sigma(C_i)\right\|\Big/ NC}{\left\|\sigma(D)\right\|},$$

$$Dis(NC) = \frac{\max_{i,j} d(c_i,c_j)}{\min_{i,j} d(c_i,c_j)}\cdot\sum\nolimits_{i=1}^{NC}\left(\sum\nolimits_{j=1}^{NC} d(c_i,c_j)\right)^{-1}.$$

Among the ten above metrics, there are three metrics (STD, RS and H) that take either separation or compactness into account. In fact, RS and H consider only separation while STD considers only compactness. All other metrics consider both separation and compactness when measuring the goodness of the clustering result.

Table 2. Datasets used for evaluation measures

#	Data set		Instance		Field	Class
	Name	Notation	Abnor.	Nom.		
1	Ionosphere	IONO	22	225	34	2
2	Wisconsin Breast Cancer	WBC	23	236	9	2
3	Wisconsin Diagnostic Breast Cancer	WDBC	35	357	30	2
4	Pima Indians Diabetes	DIAB	50	500	8	2
5	Diabetic Retinopathy Debrecen	MESS	60	606	19	2
6	Banknote Authentication	BNAU	73	738	4	2
7	Cardiotocography	CARD	164	1645	21	2
8	MAGIC gamma telescope	MAGI	1233	12332	10	2

5. EXPERIMENTAL SETTINGS

This section presents the settings of the experiments in this paper. First, the selected datasets for testing the validation metrics are introduced. After that, the data pre-processing steps are presented.

5.1. Selected datasets

The validation metrics are tested on eight benchmarking datasets drawn from UCI machine learning repository. The tested datasets, their notation and properties (pre-processed) are presented in Table 2. The detailed description of the datasets is as follows.

- Ionosphere dataset (IONO): This radar data was collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. “Good” radar returns are those showing evidence of some type of structure in the ionosphere. “Bad” returns are those that do not. The dataset includes 351 instances, each has 34 attributes.

- Wisconsin Dataset (WBC): The Wisconsin diagnostic breast cancer dataset contains 699 instances and has 9 attributes.

- The Wisconsin Diagnostic Breast Cancer (WDBC): This dataset describes nuclear characteristics for breast cancer diagnosis, also distinguishing cancer types as benign (normal) or malignant (abnormal). The dataset includes 659 samples and has 30 attributes.

- Diabetes dataset (DIAB): Several constraints were placed on the selection of these instances from a larger database. The dataset includes 768 (500 negative, 268 positive) instances and has 8 attributes.

- Diabetic Retinopathy Debrecen dataset (MESS): This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. There are 1151 instances and each instance has 19 attributes.

- Banknote Authentication dataset (BNAU): Data were extracted from images that were taken from genuine and forged banknote-like specimens. There are 1372 instances and each of them has 4 attributes.
- Cardiotocography dataset (CARD): 2126 fetal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians into a fetal state including three classes (N, S, P).
- Magic gamma telescope dataset (MAGI): The data are Monte Carlo generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. In the dataset, there are 19020 instances (12332 gamma and 6688 hadron), each has 10 attributes.

5.2. Pre-processing

Before the datasets could be used for testing the validation metrics, they need to be processed to present for the anomaly detection problems. The following pre-processing steps are applied to each dataset.

- Removing missing values: there are some methods for dealing with missing values in a dataset. In this paper, to avoid side-effect from pre-processing to the accuracy of the internal measures, the instances that have missing values are removed from datasets.
- Removing duplicated samples: the datasets may have some duplicate instances that may affect to the calculation of validation metrics. In this paper, each instance in a dataset will be checked for duplication, then all other identical ones will be removed.
- Normalization: datasets are usually normalized to avoid dominance of some properties over the others. The method is chosen to normalize datasets is as follows:

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{11}$$

  After being normalized, the value of all properties in the datasets are in range [0, 1].
- Down-sampling: To produce datasets that are suitable for the anomaly detection problem, the original datasets are down-sampled. Instances of the biggest class are retained to form normal class, while instances from others class are randomly selected to create abnormal class. The rate between number of abnormal instances and normal instance is 1/10.

6. RESULTS AND DISCUSSION

We first calculated the value of each metric when the number of abnormal samples that are swapped to the normal cluster is varied from 10% to 100%. The results of 10 validation metrics are compared with the result obtained by using logistic regression [17]. Table 3 and 4 present the comparison between ten validation metrics and logistic regression (F1) on two datasets WBC and BNAU, respectively [1]. In these tables, if a value is smaller than 1, this value is presented as $.xyz$ (for example 0.298 is presented as .298) and the first column (%) presents the number of the samples swapped between two clusters. Moreover, in each metric, if any result that does not follow the desired rule of the metric, this result is printed bold faced.

Table 3. Values of measures on dataset WBC

%	STD	RS	H	CH	I	D	S	DB	XB	SD	F1
0	.298	.152	.708	46.20	.529	.114	.388	1.02	.45	2.42	.759
10	.303	.122	.619	35.77	.408	.041	**.502**	1.26	.58	2.91	.615
20	.309	.089	.512	25.17	.285	**.041**	.469	1.64	.83	3.62	.466
30	.312	.072	.446	19.86	.225	.040	.406	1.89	1.05	4.12	.447
40	.316	.045	.345	12.14	.137	**.040**	.317	2.54	1.71	5.47	.348
50	.319	.027	.260	7.23	.081	**.040**	.221	3.35	2.88	7.15	.327
60	.320	.023	.237	6.04	.068	**.040**	.162	3.68	3.44	7.81	.263
70	.321	.017	.202	4.45	.050	**.040**	.116	4.30	4.67	9.13	.250
80	.323	.007	.125	1.78	.020	**.040**	.064	6.79	11.71	14.36	.203
90	.324	.002	.056	.39	.004	**.040**	.024	14.09	53.38	29.97	.206
100	**.323**	**.003**	**.076**	**.73**	**.008**	**.099**	.009	**10.09**	**28.40**	**21.49**	.158

It can be seen from Table 3 that most metrics produce a consistent result when the number of swapped abnormal samples varied. Apparently, there are nine out of ten metrics (excepts D) that are very good for measuring the quality of anomaly detection approaches. With these measures, there is only one result that did not follow the expected rule. Conversely, the results on metric D are not as good as the others. The metric D produces many results that did not follow the metric's rules on this dataset. Moreover, the logistic regression approach [17] produces the best result on this tested problem.

The results on Table 4 are mostly consistent with the results in Table 3. It can be observed that metric D also did not produce good results on this dataset. Moreover, the nine metrics that produced the good result in Table 3 are also the good metrics in Table 4. However, the results of metric S are not as consistent as those in the previous table. This metric created 2 inconsistent results on this dataset. For logistic regression approach, its result is still very good on this problems with only one inconsistent value.

The last result presented in the section is the Pearson correlation coefficient of each metric on the tested dataset. The Pearson correlation coefficients of ten metrics and of the logistic regression approach are shown in Table 5. In this table, the last row presents the average correlation value of each metrics over the tested datasets. Moreover, if the average value of a metric is better than the logistic regression method, its average value is printed bold faced, and if it is worse than the logistic regression method, the value is printed italic faced.

It can be seen from Table 5 that, on average, there are is only one metric (H) that is better than the logistic regression method when being used for evaluating the performance

[1] The results on other datasets are consistent with those in these tables and they are not presented for the succinct presentation of the paper.

Table 4. Values of indices with data set BNAU

%	STD	RS	H	CH	I	D	S	DB	XB	SD	F1
0	.175	.098	.082	87.57	.044	.052	.222	1.70	.76	10.34	.848
10	.177	.083	.074	72.89	.037	.011	**.293**	1.89	.91	11.42	.747
20	.178	.072	.068	63.23	.032	.010	**.295**	2.07	1.05	12.36	.630
30	.180	.054	.056	46.02	.023	.009	.257	2.43	1.44	14.48	.534
40	.181	.043	.050	36.19	.018	.006	.210	2.77	1.83	16.53	.444
50	.182	.025	.036	20.70	.011	.006	.155	3.57	3.20	21.10	.345
60	.183	.018	.029	14.61	.007	.005	.111	4.22	4.54	24.88	.252
70	.184	.011	.023	9.40	.005	**.005**	.075	5.11	7.05	30.24	.213
80	.184	.007	.017	5.49	.003	.002	.049	6.47	12.07	38.12	.183
90	.184	.001	.007	1.12	.001	.000	.024	13.80	59.36	81.53	.164
100	**.184**	**.002**	**.008**	**1.27**	**.001**	**.000**	.015	**12.69**	**52.14**	**75.40**	**.196**

Table 5. Correlation coefficient of the validation measures

Dataset	STD	RS	H	CH	I	D	S	DB	XB	SD	F1
IONO	.981	-.981	-.982	-.980	-.980	-.494	-.916	.874	.764	.869	-.955
WBC	.943	-.946	-.984	-.933	-.932	-.095	-.956	.862	.718	.858	-.948
WDBC	.932	-.937	-.972	-.915	-.923	-.846	-.947	.717	.545	.719	-.974
DIAB	.928	-.929	-.969	-.927	-.926	-.500	-.972	.914	.806	.912	-.951
MESS	.965	-.965	-.990	-.965	-.965	-.671	-.892	.924	.845	.926	-.931
BNAU	.969	-.970	-.992	-.965	-.965	-.700	-.939	.869	.761	.867	-.959
CARD	.959	-.960	-.989	-.958	-.958	-.933	-.925	.871	.750	.870	-.971
MAGI	.937	-.938	-.979	-.936	-.935	-.801	-.946	.764	.587	.765	-.976
AVG	*.952*	*-.953*	**-.982**	*-.947*	*-.948*	*-.630*	*-.937*	*.849*	*.722*	*.848*	-.958

of anomaly detection techniques. The average of the Pearson correlation coefficient of H is very close to -1 (-0.982). Therefore, this metric might be very reliable for validating the goodness of anomaly detection methods. All other metrics are worse compared to the logistic regression method. However, four metrics including STD, RS, CH and I also produce very solid results and the average value of their Pearson correlation coefficient is also very close the value of the logistic regression method. In fact, the mean of the correlation value of STD, RS, CH, and I is 0.952, -0.953, -947 and -0.948 respectively while this value for logistic regression method is 0.958. There are four out of ten tested metrics that produced substantially worse results that are: D, DB, XB, SD. Thus, these four metrics should not be used for measuring the performance of anomaly detection techniques.

One of the reason while H metric is better than other metrics in this experiment is that H metric is one of the two metrics that consider only separation. Another metric that considers only separation is RS and this metric is the second best metric amongst ten tested metrics. We suppose that for measuring the performance of anomaly detection methods, the metrics that take into account only separation are more suitable than the metrics that consider only compactness or both. In the future, we will conduct more research to investigate this hypothesis. Overall, the results in this section show that some internal validation metrics for clustering algorithms such as STD, RS and H are also very reliable for measuring for the performance of anomaly detection methods. Particularly, the modified Hubert (H) is even better than the method using logistic regression. Therefore, this method could replace the logistic regression approach when validating the goodness of unsupervised anomaly detection algorithms.

7. CONCLUSIONS AND FUTURE WORK

This paper proposed some metrics for validating the efficiency of unsupervised anomaly detection algorithms. The new metrics relies on the internal evaluation metrics used for validating the goodness of clustering algorithms. The experiments were conducted on a number of benchmarking anomaly datasets. The results showed that four of the proposed metrics produce competitive or better than the results obtained by the previous method [17]. Thus, these metrics could replace the logistic regression method [17] when being used for validating the outcome of unsupervised anomaly detection techniques.

There are some research areas for future work which arise from this paper. First, the authors would like to apply the good metrics in this paper to evaluate the performance of unsupervised anomaly detection algorithms in real world applications. In this paper, the metrics has been investigated using a number of benchmarking datasets with the available labels. In the future, the authors want to use the good metrics in this paper to evaluate the performance of various unsupervised anomaly detection algorithms when they are applied to solving real world problems such as online games cheating detection [17] and credit card fraud detection [12].

Second, the authors would like to propose better evaluation metrics for unsupervised anomaly detection algorithms. The metrics tested in this paper are often based on compactness or both compactness and separation of the clusters. This is suitable for clustering problems. However, for anomaly detection problems, it may be better if the metrics is only based on separation properties. One of such metrics is the dissimilarity measures of data in hierarchical clustering [26]. In the future, the authors would like to study this method for anomaly detection algorithms.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2014.09. Trung Thanh Nguyen was partly funded by VNG company.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, “On the importance of the pearson correlation coefficient in noise reduction,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 16, no. 4, pp. 757–765, 2008.

[2] M. Bouguessa, “A mixture model-based combination approach for outlier detection,” *International Journal on Artificial Intelligence Tools*, vol. 23, no. 4, 2014.

[3] 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, May 16-18, 2000, Dallas, Texas, USA.*, 2000, pp. 93–104.

[4] T. Caliski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[5] 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.

[6] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv.*, vol. 41, no. 3, 2009.

[7] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979.

[8] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[9] M. Gebski and R. K. Wong, “An efficient histogram method for outlier detection,” in *Advances in Databases: Concepts, Systems and Applications, 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9-12, 2007, Proceedings*, 2007, pp. 176–187.

[10] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *J. Intell. Inf. Syst.*, vol. 17, no. 2-3, pp. 107–145, 2001.

[11] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, “Quality scheme assessment in the clustering process,” in *Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings*, 2000, pp. 265–276.

[12] N. S. Halvaiee and M. K. Akbari, “A novel model for credit card fraud detection using artificial immune systems,” *Appl. Soft Comput.*, vol. 24, pp. 40–49, 2014.

[13] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[14] E. León, O. Nasraoui, and J. Gómez, “Anomaly detection based on unsupervised niche clustering with application to network intrusion detection,” in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2004, 19-23 June 2004, Portland, OR, USA*, 2004, pp. 502–508.

[15] H. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander, “On the internal evaluation of unsupervised outlier detection,” in *Proceedings of the 27th International Conference on Scientific and Statistical Database Management, SSDBM ’15, La Jolla, CA, USA, June 29 - July 1, 2015*, 2015, pp. 7:1–7:12.

[16] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002.

[17] T. T. Nguyen, A. T. Nguyen, T. A. H. Nguyen, L. T. Vu, Q. U. Nguyen, and L. D. Hai, “Unsupervised anomaly detection in online game,” in *Proceedings of the Sixth International Symposium on Information and Communication Technology, Hue City, Vietnam, December 3-4, 2015*, 2015, pp. 4–10.

[18] A. Patcha and J. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[19] G. Rätsch, S. Mika, B. Schölkopf, and K. Müller, "Constructing boosting algorithms from svms: An application to one-class classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1184–1199, 2002.

[20] V. Roth, "Kernel fisher discriminants for outlier detection," *Neural Computation*, vol. 18, no. 4, pp. 942–960, 2006.

[21] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

[22] S. Sharma, *Applied Multivariate Techniques.* New York, NY, USA: John Wiley & Sons, Inc., 1996.

[23] S. Singh and M. Markou, "An approach to novelty detection applied to the classification of image regions," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 396–407, 2004.

[24] G. J. Székely and M. L. Rizzo, "Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method," *J. Classification*, vol. 22, no. 2, pp. 151–183, 2005.

[25] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Addison-Wesley, 2005.

[26] D. Wei, Q. Jiang, Y. Wei, and S. Wang, "A novel hierarchical clustering algorithm for gene sequences," *BMC Bioinformatics*, vol. 13, p. 174, 2012.

[27] W. Wong, A. W. Moore, G. F. Cooper, and M. M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, 2003, pp. 808–815.

[28] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, 1991.

[29] D. Yu, G. Sheikholeslami, and A. Zhang, "Findout: Finding outliers in very large datasets," *Knowl. Inf. Syst.*, vol. 4, no. 4, pp. 387–412, 2002.

[30] J. Zhang and H. H. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance," *Knowl. Inf. Syst.*, vol. 10, no. 3, pp. 333–355, 2006.

[31] A. Zimek, R. J. G. B. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: challenges and research questions a position paper," *SIGKDD Explorations*, vol. 15, no. 1, pp. 11–22, 2013.

Received on June 30 - 2016
Revised on February 28 - 2017