ICIC Express Letters
Part B: Applications
Volume **7**, Number **10**, October **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2155–2160**

OUTLIER DETECTION APPROACH BASED ON LOCAL OUTLIER FACTOR FOR DATASETS WITH MIXED ATTRIBUTES

Taegu Kim[1] and Nam-Wook Cho[2,*]

[1]Department of Industrial and Management Engineering
Hanbat National University
125, Dongseodae-ro, Yuseong-gu, Daejeon 305-719, Korea
taegu.kim@hanbat.ac.kr

[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 March 2016; accepted June 2016

Abstract. *Although outlier detection has received significant attention by practitioners as well as researchers, its application to datasets consisting of both categorical and numerical attributes still remains a challenge. In this paper, a novel approach based on the local outlier factor (LOF) and similarity measure is proposed to tackle the challenge. Occurrence frequency similarity is adopted to measure the closeness of categorical data and derive a continuous distance accordingly. Two distances from categorical and numerical attributes are merged and input to the LOF calculation to identify outliers. Test results on various datasets confirm that the proposed approach provides superior performance for all cases compared to the simple numerical approach. The consistent superiority over the benchmark validates that the similarity measure successfully captures the characteristics of categorical data.*

Keywords: Outlier detection, Local outlier factor, Mixed type data, Categorical data, Similarity

1. **Introduction.** Outlier detection attempts to identify an outlying observation that deviates significantly from the other observations [1-3]. An outlier is defined as an object that deviates far from the majority so as to arouse suspicion that it was generated by a different mechanism [4].

Density-based approaches identify outliers as those lying in low-density areas. The local outlier factor algorithm is a density-based algorithm that has been widely used in many applications [5]. The method identifies local outliers based on the local density of an object's neighborhood, defined as the local outlier factor (LOF) [6,7]. The LOF approach has advantages over other outlier detection techniques. It is not influenced by the distribution of normal behavior [5]. Also, it has been demonstrated by Lazarevic et al. [8] that LOF typically achieved better performance in network intrusion identification than existing outlier detection algorithms.

Despite these advantages of the LOF method, its application often has a limitation regarding the data type. The technique assumes the input data to be numerical; however, real data typically contain both numerical and categorical attributes [9-11]. Therefore, as Wei et al. [9] and Zhao et al. [12] reported, a density-based outlier detection method cannot effectively address categorical data. Consequently, for the LOF method, the problem of identifying outliers in categorical data is more challenging than the problem of identifying dataset outliers in numerical data [13].

This research proposes an LOF-based local outlier detection method for a dataset including categorical and numerical attributes. To address the categorical attributes of a

dataset, the similarity concept is utilized. Similarity is a numerical measure of the resemblance between two categorical instances and allows the calculation of the Euclidean distance between them. Then, the two distances from the categorical and numerical attributes are merged and the LOF method is then applied to the resultant distance. To verify the practical usefulness of the proposed approach, a set of experiments on real datasets including categorical and numerical attributes was conducted.

2. **Methodology.**

2.1. **Similarity measure.** To address the categorical attributes of a dataset, the similarity concept is utilized. In this paper, occurrence frequency measure (OF) is selected as a similarity measure, since it can incorporate the frequency of the categorical attributes. According to the OF measure, the similarity of two identical data records always is one, whereas the similarity of mismatching cases is determined by their frequencies. The OF similarity between two instances X and Y can be calculated with following function.

$$\begin{aligned} S(X,Y) &= \sum_{k=1}^{d} w_k S_k(X_k, Y_k) \\ S_k(X_k, Y_k) &= \begin{cases} 1 & \text{if } X_k = Y_k \\ \left[1 + \ln \frac{N}{f_k(X_k)} \ln \frac{N}{f_k(Y_k)}\right]^{-1} & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

$f_k(X_k)$ is the number of times that the value X_k appears for the attribute k where N is the size of the dataset. For example, if an instance X is a person and the attribute k is a location, X_k is the location of person X and $f_k(X_k)$ is the number of individuals located at X_k. Therefore, if he or she lives in an isolated location, $f_k(X_k)$ and $S_k(X_k, Y_k)$ become small.

$S_k(X_k, Y_k)$ is the per attribute similarity and the total similarity $S(X, Y)$ can be derived as a weighted sum of all the $S_k(X_k, Y_k)$. For the OF methodology, the weight coefficients w_k are identical to $1/d$ for all attributes $k = 1, 2, \ldots, d$.

2.2. **Distance.** The Euclidean distance between two instances with numerical attributes is defined as follows:

$$dist(X,Y) = \sum_{k=1}^{d} (x_k - y_k)^2 \tag{2}$$

where x_k is the value of instance X regarding attribute k.

A distance based on categorical values can be computed using their similarity.

$$dist(X,Y) = \frac{1}{S(X,Y)} - 1 \tag{3}$$

This equation describes the relation between two measures. If the values of two categorical instances are the same for all attributes, their similarity is one and thus, the distance becomes zero. This coincides with the Euclidean distance of two overlapped points; otherwise, the similarity is less than one and the distance increases. If the similarity becomes considerably smaller because one of the instances includes unusual values, it can be interpreted that the instance is located distant from the others in the manner of Euclidean space.

Because mixed type data consisting of numerical and categorical attributes is our research interest, it is necessary to merge the different distance measures. One possible approach is a simple sum. However, distances should first be normalized because their scales could be different.

$$\begin{aligned} d &= d_N + d_C \\ d &= \bar{d}_N + \bar{d}_C \end{aligned} \tag{4}$$

where $d_N(d_C)$ is a numerical (categorical) distance and $\bar{d}$ denotes a normalized distance. Another approach is to utilize the definition of numerical distance as follows:

$$d = \sqrt{d_N^2 + d_C^2} \tag{5}$$

Further, the size of the dimension, i.e., the number of attributes, can also be considered as a weight. This is presumed to reflect the relative amount of information from the two sources. In consideration of these factors, two merging schemes are suggested in the following table.

Table 1. Schemes for merging distances

Index	Description	Equation
a	Simple, normalized, weighted	$d = \sqrt{w_N}\bar{d}_N + \sqrt{w_C}\bar{d}_C$
b	Squared, normalized, weighted	$d = \sqrt{w_N\bar{d}_N^2 + w_C\bar{d}_C^2}$

w_N and w_C are the number of numerical and categorical attributes in the data

3. **Experiment and Results.** In this section, a set of experiments was performed to examine the effectiveness of the proposed method.

3.1. **Data.** The data used in this research were provided by the UCI Machine Learning Repository [14]. Among the datasets, Adult, German Credit, Bike Sharing, Solar Flare, and Yeast datasets were used in the experiment.

3.2. **Experiment design.** For each dataset, the test procedure consisted of preparation and calculation stages.

In the preparation stage, k-means clustering has been applied to the datasets. Among the groups defined by the k-means clustering, the group with the fewest instances was assumed to be the outliers. The test data is composed of 5% outliers and 95% normal instances. After the sampling, the categorical variables were reassigned with random integers.

In the first step of the calculation stage, two kinds of distances were calculated for the numerical and categorical variables. For the numerical variables, the distance was easily derived after a simple normalization to the zero-one interval for each variable. We followed the OF similarity method to measure the categorical distance. Subsequently, these distances were combined with two different merging schemes as follows:

$$\begin{aligned} &\text{a.} \quad d = \sqrt{\bar{d}_N^2 + \bar{d}_C^2} \\ &\text{b.} \quad d = \sqrt{w_N\bar{d}_N^2 + w_C\bar{d}_C^2} \end{aligned} \tag{6}$$

In both equations, the weight coefficients w_N and w_C denote the number of numerical and categorical variables and $\bar{d}_N$ and $\bar{d}_C$ are the normalized numerical and categorical distances, respectively.

Based on the total distance derived in Equation (6), the LOF for each instance in the sample data was determined with the assumed number of neighbors, *minpts* (5, 10, 15, or 20). Sorting the instances into their descending orders of LOF values, we selected a set of candidates for outliers. The sizes of the candidate sets varied from one to ten percent of the total test data. For example, the test data sampled from the Adult dataset consisted of 1,000 instances and thus cases of the size of n were one of 10, 20, ..., 100.

Finally, the performance of the combination $(n, minpts)$ was evaluated considering recall and precision. Therefore, 160 results were gathered for each raw data: combination of two weighting schemes, four neighbor sizes, ten candidate sizes, and two performance

measures. Furthermore, as a benchmark or baseline method, we prepared a control group result using a simple numerical distance method that considered all variables as simply numerical.

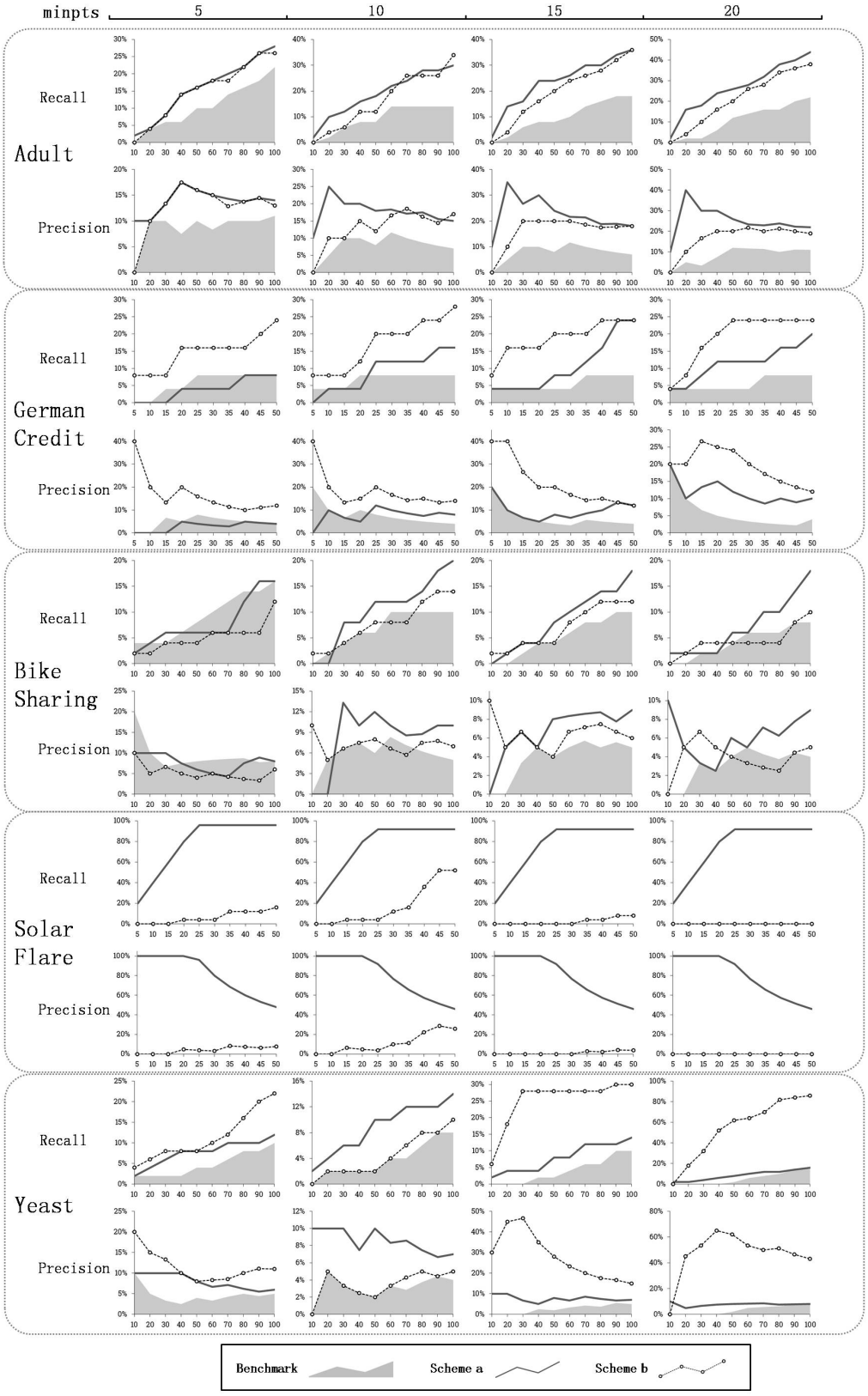

FIGURE 1. Test results

3.3. **Results.** Figure 1 presents the results for the five datasets considering recall and precision. The dark gray solid line indicates the results from the first weighting scheme, Equation (6)a (scheme a) and the black dashed line connecting circles corresponds to the other, Equation (6)b (scheme b). The benchmark is displayed as gray color shades. Overall, the suggested detecting method exhibited superior performance, both in recall and precision, compared to the benchmark regardless of the dataset, number of neighbors, size of candidate set, or weighting scheme. However, there was not a decisive difference between the weighting schemes.

A statistical analysis of the performance is summarized in Table 2. For each dataset, a paired t-test was conducted. The third and fourth columns display the comparison results between the method used and the benchmark, whereas the difference between the two weighting schemes is presented in the last column. Their performance did not move together. Instead, one scheme was significantly better than the other depending on the data. In summary, the proposed method was more effectively managing mixed type data. It was proven that the similarity measure was remarkably beneficial because of the consistent superiority regardless of the data and weighting schemes.

Table 2. Performance comparison

Data	Measure	Scheme a vs. benchmark	Scheme b vs. benchmark	Scheme a vs. scheme b
Adult	Recall	**scheme a*****	**scheme b*****	**scheme a*****
	Precision	**scheme a*****	**scheme b*****	**scheme a*****
German Credit	Recall	**scheme a*****	**scheme b*****	**scheme b*****
	Precision	**scheme a****	**scheme b*****	**scheme b*****
Bike Sharing	Recall	**scheme a*****	**benchmark**	**scheme a*****
	Precision	**scheme a*****	**scheme b**	**scheme a****
Solar Flare	Recall	**scheme a*****	**scheme b*****	**scheme a*****
	Precision	**scheme a*****	**scheme b*****	**scheme a*****
Yeast	Recall	**scheme a*****	**scheme b*****	**scheme b*****
	Precision	**scheme a*****	**scheme b*****	**scheme b*****

Each cell indicates the superior model using the result of a paired t-test
Statistical significance: $^{***}1\%$, $^{**}5\%$, and $^{*}10\%$

4. **Conclusion.** In this paper, an LOF-based local outlier detection method for a dataset mixed with categorical and numerical attributes was proposed. A similarity measure was used to determine the distance between categorical attributes, whereas the Euclidean distance measure was used for numerical attributes. Then, the two distance measures were merged to determine a distance between the data points. As presented in the set of experiments, the proposed method can be applied effectively to various real cases with both categorical and numerical attributes using the advantages of the LOF method. Despite the advantages of the proposed method, its application to a larger-scale dataset can be limited owing to its time complexity. This is frequently the case with the LOF algorithm. Therefore, reducing the time complexity of the LOF algorithm is a suitable topic for future research.

Acknowledgement. This work was supported by the research program funded by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST) (NRF-2013R1A1A22011169).

REFERENCES

[1] 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, pp.9587-9596, 2011.

[2] Z. He, S. Deng and X. Xu, An optimization model for outlier detection in categorical data, *Advances in Intelligent Computing*, vol.3644, pp.400-409, 2005.

[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*, pp.210-217, 2007.

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

[5] D. Pokrajac, A. Lazarevic and L. J. Latecki, Incremental local outlier detection for data streams, *IEEE Symposium on Computational Intelligence and Data Mining*, pp.504-515, 2007.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, LOF: Identifying density-based local outliers, *SIGMOD Rec.*, vol.29, pp.93-104, 2000.

[7] M. A. Maloof, *Machine Learning and Data Mining for Computer Security: Methods and Applications (Advanced Information and Knowledge Processing)*, 2005.

[8] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur and J. Srivastava, A comparative study of anomaly detection schemes in network intrusion detection, *SDM*, pp.25-36, 2003.

[9] L. Wei, W. Qian, A. Zhou, W. Jin and J. Yu, HOT: Hypergraph-based outlier test for categorical data, *Advances in Knowledge Discovery and Data Mining*, vol.2637, pp.399-410, 2003.

[10] H. L. Chen, M. S. Chen and S. C. Lin, Catching the trend: A framework for clustering concept-drifting categorical data, *IEEE Trans. Knowledge and Data Engineering*, vol.21, pp.652-665, 2009.

[11] D. Ienco, R. G. Pensa and R. Meo, A semisupervised approach to the detection and characterization of outliers in categorical data, *IEEE Trans. Neural Networks and Learning Systems*, 2016.

[12] X. Zhao, J. Liang and F. Cao, A simple and effective outlier detection algorithm for categorical data, *International Journal of Machine Learning and Cybernetics*, vol.5, pp.469-477, 2014.

[13] A. Taha and A. S. Hadi, A general approach for automating outliers identification in categorical data, *2013 ACS International Conference on Computer Systems and Applications*, pp.1-8, 2013.

[14] K. Bache and M. Lichman, *UCI Machine Learning Repository*, http://archive.ics.uci.edu/ml.