A Survey on Mixed-Attribute Outlier Detection Methods

Nur Rokhman
Department of Computer Sciences and Electronics, Universitas Gadjah Mada
Daerah Istimewa Yogyakarta 55281, Indonesia
Email: nurrokhman@ugm.ac.id

***Abstract*—In the data era, outlier detection methods play an important role. The existence of outliers can provide clues to the discovery of new things, irregularities in a system, or illegal intruders. Based on the data, outlier detection methods can be classified into numerical, categorical, or mixed-attribute data. However, the study of the outlier detection methods is generally conducted for numerical data. Meanwhile, many real-life facts are presented in mixed-attribute data. In this paper, the researcher presents a survey of outlier detection methods for mixed-attribute data. The methods are classified into four types, namely, categorized, enumerated, combined, and mixed outlier detection methods for mixed-attribute data. Through this classification, the methods can be easily analyzed and improved by applying appropriate functions.**

***Index Terms*—Outlier Detection, Categorical Data, Numerical Data, Mixed-Attribute Data**

I. Introduction

OUTLIERS are data that deviates from normal data in general. Currently, the issue of outliers is important. The existence of outliers can provide clues to the discovery of new things, irregularities in a system, or illegal intruders. Outlier detection methods can be classified into three categories, namely, supervised, semi-supervised, and unsupervised methods. The supervised outlier detection method is characterized by the presence of a trained dataset that models normal and abnormal conditions. In this method, outlier detection is conducted by identifying the existence of data, whether it is in the normal or abnormal dataset. Several examples of outlier detection methods in this category include artificial neural networks, statistics, and rule-based models.

Next, the semi-supervised outlier detection method is characterized by the presence of trained datasets for normal conditions. However, there is no trained dataset for abnormal conditions. In this method, outlier detection is conducted by identifying the existence of data in the set of normal conditions. Several examples of outlier detection methods in this category include support vector machine and hidden Markov model.

The unsupervised outlier detection method does not have any trained dataset. In this method, outlier identification is conducted by comparing each data point to the others, which will yield the outlier degree of the data point. Then, the outlier degree is compared to a threshold value.

In real life, facts are presented not only in numerical data but also in categorical and mixed-attribute data. Thus, mixed-attribute data is a mixture of numerical and categorical data. Based on the data to be processed, outlier detection methods can be classified into three types, namely, the outlier detection method for numerical, categorical, and mixed-attribute data. However, generally, the developed outlier detection methods only work for numerical data. Many surveys such as those presented in Refs. [1–9] have been conducted to classify the outlier detection methods for numerical data.

II. Research Method

A. Outlier Detection Methods for Mixed-Attribute Data

In various cases, categorical and numerical data may appear simultaneously in a dataset which called mixed-attribute data [10]. A data point on the mixed-attribute dataset can be modeled with $P = (P_c, P_q)$ where P_c denotes the categorical data and P_q denotes the numerical data [11]. Based on the construction of outlier value, the outlier detection methods for mixed-attribute data can be classified into four types, namely categorized, enumerated, combined, and mixed outlier detection methods.

Then, f is a categorize function that converts numerical data into categorical data, g is an enumerate function that converts categorical data into numerical data, and h is a real value function. Next, the researcher defines ODC as a function for determining the outlier

Received: Mar. 23, 2019; received in revised form: Jun. 01, 2019; accepted: Jun. 02, 2019; available online: Jun. 04, 2019.

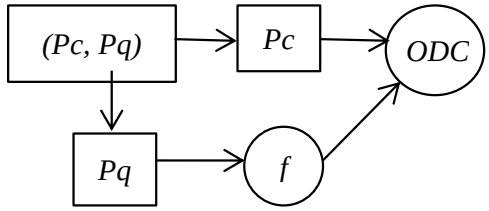

Fig. 1. Categorized outlier detection method for mixed-attribute data.

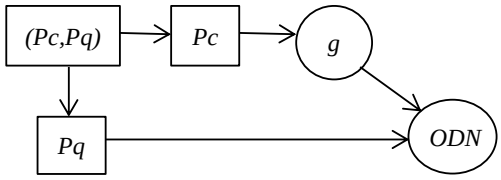

Fig. 2. Enumerated outlier detection method for mixed-attribute.

value of categorical data, and ODN for numerical data. The process of the categorized outlier detection method for mixed-attribute data is started by separating the numerical data (P_q) from the categorical data (P_c). Then, P_q is converted into categorical data using f. The outlier value (OV) is determined using ODC, as presented in Eq. (1). This process is depicted in Fig. 1.

$$\text{OV} = \text{ODC}(P_c, f(P_q)). \tag{1}$$

The process of the enumerated outlier detection method for mixed-attribute data starts from separating the categorical data (P_c) from the numerical data (P_q). Then, P_c is converted into numerical data using g. The outlier value is determined using ODN, as presented in Eq. (2). This process is depicted in Fig. 2.

$$\text{OV} = \text{ODC}(P_q, g(P_c)). \tag{2}$$

The process of combined outlier detection method for mixed-attribute data is from separating the categorical data (P_c) from the numerical data (P_q). Then, ODC is used to calculate the outlier value of P_c. Then, ODN is used to calculate the outlier value of P_q. It combines the results of ODC and ODN using h yields the final outlier value as presented in Eq. (3). This process is shown in Fig. 3.

$$\text{OV} = h(\text{ODC}(P_c, \text{ODN}(P_q)). \tag{3}$$

The mixed outlier detection method for mixed-attribute data considers the relationship between categorical data (P_c) and numerical data (P_q). The process starts with extracting the numerical properties of the data. Then, the relationships between data and the numerical properties of data are used to determine

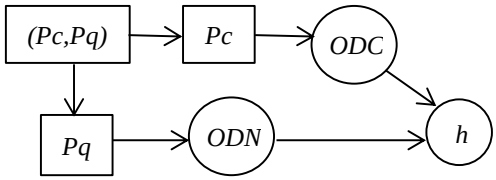

Fig. 3. Combined outlier detection method for mixed-attribute data.

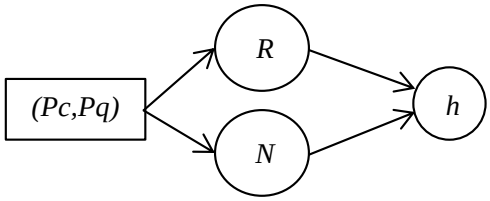

Fig. 4. Mixed outlier detection method for mixed-attribute data.

the outlier value of the entire data by using h. Then, R is a function used to determine the relationship between data, and N is a function used to determine the numerical property of data. The outlier detection method can be expressed in Eq. (4). This process is shown in Fig. 4.

$$\text{OV} = h(N(Pc, Pq), R(P_c, Pq)). \tag{4}$$

III. Results and Discussion

Figures 1–4 show the different classes of outlier detection methods for mixed-attribute data. The differences lie in the transformation. In this section, the outlier detection methods in each class are discussed.

A. Categorized Outlier Detection Method for Mixed-Attribute Data

As shown in Fig. 1, in this method, numerical data are transformed into categorical data. Then, outlier detection is conducted using the outlier detection method for categorical data. Various methods in this class have been developed. These methods are characterized by the function to convert numerical data into categorical data and the categorical outlier detection method.

Numerical data are transformed into categorical data by discretization like the dataset contains only categorical attributes. Then, the analysis is focused on the association relationships between data. In this method, the saliency degree is used to measure the outlier value. It is evaluated from the values of associated attributes [12].

In segmentation-and-combination-based detection of anomalies, continuous or numerical data are converted into categorical data by segmenting the data using a

TABLE I
CATEGORIZED OUTLIER DETECTION METHODS FOR MIXED-ATTRIBUTE DATA.

Refs.	Categorize functions	Outlier detection methods
[12]	Discretization	This method transforms numerical data into categorical data by discretization. The outlier value is determined by using the saliency degree, which is calculated from the relationships between data.
[13]	Interval	Numerical data are transformed into categorical data using a specific interval length. The transformation is repeated using different interval lengths. The outlier is extreme value data which have low-frequency on each repetition.
[14]	z-discretization and k-means clustering	The numerical data are discretized by using z-discretization and k-means clustering. The outlier detection is carried out by using AVF method.

specific interval length. The segmentation is repeated using different interval lengths. On each repetition, the low-frequency data are identified. The outliers are extreme value data which have low frequency on each repetition [13].

Numerical data are transformed into outlier data using z-discretization and k-means clustering. Outlier detection is conducted using Attribute Value Frequency (AVF) method [14]. All of these methods differ from the function used to transform numerical data into categorical data, and the function used to detect the outlier. Table I shows a comparison of these methods. These methods may be improved by selecting the appropriate function used to convert numerical data into categorical data. The function is used to conduct the categorical outlier detection.

Comparing the three methods shows the fact that descretization technique is the fastest methods. The interval technique needs to repeat the processes. This makes a higher complexity level.

B. Enumerated Outlier Detection Method for Mixed-Attribute Data

As shown in Fig. 2, categorical data are transformed into numerical data. Then, outlier detection is conducted by using the outlier detection method for numerical data. Various methods in this class have been developed. These methods are characterized by the function used to convert categorical data into numerical data and the categorical outlier detection method.

Reducing memory is used to solve the high memory consumption problem in Link-based Outlier and Anomaly Detection in Evolving Datasets (LOADED). In this method, categorical data are transformed into numerical data using Naive Bayes classifier. Then, the outlier value is calculated using a covariance matrix for the entire data [15].

Buff detection method considers the mutual correlation between data. In this method, the attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. Each attribute is mapped to a corresponding latent numerical variable via a specific Link function technique, such as logit function for binary attributes and log function for count attributes. The dependency between mixed-type attributes is expressed by the relationship between their latent variables using a variance-covariance matrix. The outlier is detected by fitting the data into the model [16]. Both methods differ on the function used to transform categorical data into numerical data, and the function used to detect the outlier. Table II shows a comparison of these methods.

By comparing both methods in Table II, it shows the fact that developing the model in the Link function technique needs many complex steps. Thus, the Naive Bayes classifier techniques have a better complexity. These methods may be improved by selecting the appropriate function to convert the categorical data into numerical data and the function to conduct numerical outlier detection.

C. Combined Outlier Detection Method for Mixed-Attribute Data

As shown in Fig. 3, the outlier detection process is conducted to categorical dan numerical data separately. Then, a function is used to combined the outlier scores, which results in the entire outlier score. The discussion about mixed-attribute outlier detection methods starts with LOADED. In this method, a similarity metric is used to process the categorical data and correlation matrix for the numerical data. LOADED uses the matrix to express the dependence between categorical and numerical data. The outlier value is calculated using data dependence [11].

In Outlier Detection for Mixed-Attribute Datasets (ODMAD), the outlier value of categorical data is calculated from its frequency. ODMAD assumes that data points that share a categorical value should also share similar continuous value. Therefore, ODMAD restricts the search space by focusing on points that share a categorical value and ranks these points based on their similarity to each other. The similarity of numerical data is calculated using the cosine similarity [17]. The lower bound and the upper bound of the similarity score is used to determine the outlierness of data.

TABLE II
ENUMERATED OUTLIER DETECTION METHODS FOR MIXED-ATTRIBUTE DATA.

Refs.	Enumerate functions	Outlier detection methods
[15]	Naive Bayes classifier	Categorical data are transformed into numerical data using Naive Bayes classifier. The outlier value is calculated using a covariance matrix.
[16]	Link function	Mixed-type attributes are mapped to latent numerical random variables that are multivariate Gaussian in nature. The outlier is detected by fitting the data into a model.

TABLE III
COMBINED OUTLIER DETECTION METHODS FOR MIXED-ATTRIBUTE DATA.

Refs.	Functions	Outlier detection methods
[11]	Similarity metric, correlation matrix	A similarity metric is used to detect the outlier in categorical data. Correlation matrix that stores Pearson's correlation coefficient between every pair of continuous attributes is used to detect the outlier in numerical data.
[17]	Frequency, cosine similarities	The outlier value of categorical data is calculated from its frequency. The similarity of numerical data is calculated using cosine similarities. The lower bound and the upper bound of the similarity score is used to determine the outlierness of data.
[18]	Decision tree, Gaussian mixture model, simple weighted linear sum	The decision tree is used to determine the outlier value of the categorical data. The Gaussian mixture model is used to determine the outlier value of the numerical data. The final outlier score is determined using a simple weighted linear sum from the outlier values.
[19]	Frequent pattern, cosine, average	The outlier value of categorical data is determined using the frequent pattern. The cosine function calculates the outlier value of the numerical data. The final outlier score is the average categorical and numerical data outlier score.
[20]	Holoentropy, HilOut algorithm	The outlier value of categorical data is determined using holoentropy. The outlier value of numerical data is determined using the HilOut algorithm. The two degrees of outlier are combined to obtain the final outlier score.
[21]	Frequency, nearest neighbor	The outlier value of categorical data is determined using data frequency. The outlier value of numerical data uses the nearest neighbor concept. Both outlier values are combined using a vector to obtain the overall outlier value.
[22]	Frequency, modified Canberra equation	The outlier value of categorical data is calculated using frequency. The outlier value of numerical data is calculated using the modified Canberra equation. Then, the average of both outlier values is the final outlier value.
[23]	Generalized linear model, robust error buffering, Gaussian predictive model	Non-numerical data are processed using the generalized linear model and robust error buffering. Numerical data are processed using the Gaussian predictive model.

According to Ref. [18], the decision tree is used to determine the outlier value of categorical data. Meanwhile, the Gaussian mixture model is used to determine the outlier value of numerical data. The final outlier score is determined using a simple weighted linear sum from the outlier values [18].

Generally, all datasets in the database do not need to be scanned. Therefore, obtaining the average categorical data outlier score with numerical data outlier scores is sufficient to determine the final score. With this assumption, the mixed-attribute outlier factor method is proposed. The outlier value of categorical data is determined using the frequent pattern. The cosine function is used to calculate the outlier factor of numerical data [19].

Categorical and numerical data are separated. Then, the outlier value of categorical data is determined using holoentropy. Then, the outlier value of numerical data is determined using the HilOut algorithm. The two degrees of outlier are combined to obtain the final outlier score [20].

Categorical and numerical data are separated. Each outlier value is determined separately. The outlier value of numerical data is determined using the nearest neighbor concept. The outlier value of categorical data is determined by using data frequency. Then, both outlier values are combined using a vector to obtain the overall outlier value [21].

Moreover, categorical and numerical data are separated. The outlier value of categorical data is calculated by using the frequency. Meanwhile, the outlier value of numerical data is calculated using the modified Canberra equation. Then, the average of both outlier values is used to determine the final outlier value [22].

Mixed-type robust detection separates the categorical and numerical data. In this method, numerical data are processed using the Gaussian predictive model. However, the non-numerical data are processed using the generalized linear model and robust error buffering [23]. Table III shows a comparison of these methods.

Comparing the methods in Table III, despite its performance, the simplest method is the combination of frequency and modified Canberra equation. These methods differ on the function to detect the outlier of categorical data, numerical data, and both outlier values. These methods may be improved by selecting the appropriate function to detect the outlier in categorical

data, the numerical data, and both outlier values.

D. Mixed Outlier Detection Method for Mixed-Attribute Data

As shown in Fig. 4, categorical and numerical data are processed simultaneously. The extracted numerical properties of datasets are used to determine the outlier value. Pattern-based Outlier Detection (POD) is proposed based on the fact that an outlier does not comply with the data pattern. In mixed attribute data, patterns are from numerical data, categorical data, and the interaction between numerical and categorical data. The more an object deviates from these patterns, the higher is its outlier factor [24]. POD uses logistic regression to learn patterns and formulate the outlier factor in mixed-attribute datasets.

IV. Conclusion

The outlier detection methods for mixed-attribute data can be classified into four groups, namely, categorized, enumerated, combined, and mixed outlier detection method. The enumerated and categorized methods transform the mixed-attribute data into single-type data. Meanwhile, the combined method merges the outlier values of categorical and numerical data. Then, the mixed method calculates the outlier value from the whole data.

The classification of outlier detection methods of mixed-type data is shown in Tables I–III. These tables clearly show that the various outlier detection methods differ on the transforming function and the outlier value determination for categorical and numerical data. Future works can be conducted to find a better method to detect outlier on mixed-attribute data by applying appropriate functions into the most efficient process, that is the minimum iteration.

References

[1] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, "A comparative study for outlier detection techniques in data mining," in *2006 IEEE Conference on Cybernetics and Intelligent Systems*. Bangkok, Thailand: IEEE, June 7–9, 2006, pp. 1–6.

[2] J. Xi, "Outlier detection algorithms in data mining," in *2008 Second International Symposium on Intelligent Information Technology Application*, vol. 1. Shanghai, China: IEEE, Dec. 20–22, 2008, pp. 94–97.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–72, 2009.

[4] P. Gogoi, D. Bhattacharyya, B. Borah, and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," *The Computer Journal*, vol. 54, no. 4, pp. 570–588, 2011.

[5] K. Singh and S. Upadhyaya, "Outlier detection: Applications and techniques," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 1, pp. 307–323, 2012.

[6] V. Ilango, R. Subramanian, and V. Vasudevan, "A five step procedure for outlier analysis in data mining," *European Journal of Scientific Research*, vol. 75, no. 3, pp. 327–339, 2012.

[7] P. Ajitha and E. Chandra, "A survey on outliers detection in distributed data mining for big data," *Journal of Basic and Applied Scentific Research*, vol. 5, no. 2, pp. 31–38, 2015.

[8] P. S. Femi and S. G. Vaidyanathan, "Comparative study of outlier detection approaches," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. Tamil Nadu, India: IEEE, July 11–12, 2018, pp. 366–371.

[9] X. Xu, H. Liu, and M. Yao, "Recent progress of anomaly detection," *Complexity*, vol. 2009, pp. 1–11, 2019.

[10] C. C. Aggarwal, *Outlier Analysis*. Switzerland: Springer International Publishing AG, 2017.

[11] A. Ghoting, M. E. Otey, and S. Parthasarathy, "LOADED: Link-based outlier and anomaly detection in evolving data sets," in *Fourth IEEE International Conference on Data Mining (ICDM'04)*. Brighton, UK: IEEE, Nov. 1–4, 2004, pp. 387–390.

[12] Y.-G. Kim and K. M. Lee, "Association-based outlier detection for mixed data," *Indian Journal of Science and Technology*, vol. 8, no. 25, pp. 1–6, 2015.

[13] R. Foorthuis, "SECODA: Segmentation-and combination-based detection of anomalies," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Tokyo, Japan: IEEE, Oct. 19–21, 2017, pp. 755–764.

[14] D. Maryono, P. Hatta, and R. Ariyuana, "Implementation of numerical attribute discretization for outlier detection on mixed attribute dataset," in *2018 International Conference on Information and Communications Technology (ICOIACT)*. Yogyakarta, Indonesia: IEEE, March 6–7, 2018, pp. 715–718.

[15] M. E. Otey, S. Parthasarathy, and A. Ghoting, "Fast lightweight outlier detection in mixed-attribute data," The Ohio State University, Tech.

Rep., 2005.

[16] Y. C. Lu, F. Chen, Y. Chen, and C. T. Lu, "A generalized student-t based approach to mixed-type anomaly detection," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, Bellevue, Washington, USA, July 14–18, 2013.

[17] A. Koufakou, M. Georgiopoulos, and G. Anagnostopoulos, "Detecting outliers in high-dimensional datasets with mixed attributes," in *Proceedings of The 2008 International Conference on Data Mining, DMIN 2008*. Las Vegas, USA: CSREA Press, July 14–17, 2008.

[18] K. N. Tran and H. Jin, "Detecting network anomalies in mixed-attribute data sets," in *2010 Third International Conference on Knowledge Discovery and Data Mining*. Phuket, Thailand: IEEE, Jan. 9–10 2010, pp. 383–386.

[19] M. K. Murthy, A. Govardhan, and L. D. SreenivasaReddy, "A model to find outliers in mixed-attribute datasets using mixed attribute outlier factor," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 5, pp. 215–219, 2013.

[20] G. A. Jahanban and T. S. Singh, "Detection of outlier schema for mixed data using ITB-SP and HilOut algorithms," *International Research Journal of Engineering and Technology (IRJET)*, vol. 01, no. 01, pp. 20–22, 2014.

[21] M. Bouguessa, "A practical outlier detection approach for mixed-attribute data," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8637–8649, 2015.

[22] B. A. Manjunatha and P. Gogoi, "Anomaly based intrusion detection in mixed attribute dataset using data mining methods," *Journal of Artificial Intelligence*, vol. 9, no. 1–3, pp. 1–11, 2016.

[23] Y. C. Lu, F. Chen, Y. Wang, and C. T. Lu, "Discovering anomalies on mixed-type data using a generalized student-t based approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2582–2595, 2016.

[24] K. Zhang and H. Jin, "An effective pattern based outlier detection approach for mixed attribute data," in *Australasian Joint Conference on Artificial Intelligence*. Adelaide, Australia: Springer, Dec. 7–10, 2010, pp. 122–131.