Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING THE PERFORMANCE OF OUTLIER DETECTION METHODS FOR CATEGORICAL DATA BY USING WEIGHTING FUNCTION

[1]NUR ROKHMAN, [2]SUBANAR, [3]EDI WINARKO
[1]Gadjah Mada University, Department of Computer Science and Electronics, Yogyakarta, Indonesia
[2] Gadjah Mada University, Department of Mathematics, Yogyakarta, Indonesia
[3] Gadjah Mada University, Department of Computer Science and Electronics, Yogyakarta, Indonesia

E-mail: [1]nurrokhman@ugm.ac.id, [2]subanar@yahoo.com, [3]ewinarko@ugm.ac.id

ABSTRACT

Outliers are uncommon events in real life. For a database processing, outlier means unusual records comparing to the remaining records. An outlier can be caused by a damage of a system. A new fact in a system can also cause outliers. Outlier detection is an important task to find an exceptional data.

Outlier detection methods for categorical data such as AVF, MR-AVF, AEVF, NAVF, OPAVF, WDOD, and FuzzyAVF work base on the attribute value data frequency. These methods start the outlier detection process by calculating the data of attribute value frequency on each attribute. Then, many complicated calculations based on the various mathematical background are carried out to find the outlier by using the data of attribute value frequency.

All the methods above do not take into account on the sparseness of each attribute. In this paper, weighting functions is used to take into account the sparseness of each attribute. AVF and WDOD methods are modified by using weighting function. The performance of these modified methods is observed based on the detected outlier of UCI Machine Learning datasets. The experiment shows that weighting function can improve the performance of AVF and WDOD on the outlier detection in Adult, Mushroom, and Nursery datasets.

Keywords: *Outlier Detection, Categorical Data, Weighting Function.*

1. INTRODUCTION

Outlier detection is an important step in data processing. Outlier detection is used to find the uncommon data. Uncommon data may be caused by illegal intrusion, a damage of system, fraud, or medical problem. It is stated by [1] that outlying observation or outlier is one that appears to deviate markedly from other members of the sample in which it occurs. An outlier might be generated by a different mechanism of the systems [2] and very low frequency [3]. Outliers or anomalies are patterns in data that do not conform to a well define notion of normal behavior [4].

Many outlier detection methods have been developed. Most of the existing methods are focused on processing numerical data. Statistically based method, distance based method, and density-based method are the common methods for numerical data.

For non-numerical data, a mapping process to numerical value is needed. AVF uses frequency data as the numerical value such that the outlier data can be determined [5]. Similarity-dissimilarity concept with contingency table to is used to determine the graphical plot of categorical data [6]. This numerical value can be used to find the outlier. Categorical data is converted into numerical data by using co-occurrence theory, which explores the relationship among items to define the similarity between pairs of objects [7]. WDOD uses attribute value frequency and average density to detect outlier of categorical data [8]. A complete evaluation on various mechanisms which maps categorical data into numerical data such that outlier can be detected has been done [9].

All of those outlier detection methods above do not take account on the sparseness of the attribute value. This paper discusses the sparseness of the attribute value, its contribution to the outlier level., and its usage on the outlier detection algorithm.

This paper examines weighting functions as a preprocessing step for the attribute value frequency

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

before the complex calculations are carried out. Weighting function makes data with the more sparse attribute value frequecy be the most possible outlier. The performance of the weighting functions is observed by their effect on the capability of finding the outlier data.

In this research, two new algorithms called WAVF and WADOD are designed. The effectivenesses of these new methods are observed by comparing its performance to the performance of AVF and WDOD. The performance of an algorithm is determined by the amount of the outlier data which is detected by the algorithm. Three datasets from UCI Machine Learning repository, namely *Mushroom*, *Nursery*, and *Adult* are used as the case study.

The organization of the remaining paper is as follows: Section 2 presents literate review. In Section 3, we describe the proposed algorithm. Section 4 contains the experimental setup, results, and discussions. Section 5 summarizes the conclusion and future work.

2. PREVIOUS WORK

There are three categories of outlier detection method, namely: supervised, semi-supervised, and unsupervised outlier detection methods. For the supervised outlier detection method, the class of normal data and the class of outlier data are available. Supervised outlier detection method is like a predictive data. Semi-supervised outlier detection method only has a class of normal data. Outliers are data which do not belong to the class of normal model. The unsupervised outlier detection method does not have any class of normal data nor class of outlier data. Unsupervised outlier detection method makes implicit assumption that: (1) normal instances are much more frequent than the outlier instances, (2) the outlier instances lie far from the normal instances, (3) normal instances are much more dense than the outlier instances [4].

For the categorical data, the existing methods such as: AVF [5], MR-AVF [10], AEVF [11, NAVF [12], OPAVF [13], FuzzyAVF [14], and WDOD [8] work base on frequency of the attribute value. WDOD uses attribute value frequency and average data density to detect the outlier [8]. These methods belong to the unsupervised outlier detection methods.

All of those methods above do not take into account data sparseness on each attribute. In Table 2, The instances x_1 and x_3 have attribute value frequency {2, 1, 3} and {1, 2, 3} respectively. By using AVF method, x_1 and x_3 have the same outlier score, even though they have different combination of attribute value frequency. On the other hand, each attribute has different spareness. These values have the same effect to the outlier score since the sparseness of the attributes is not taken into account. Finally, the instances x_1 and x_3 have the same AVF score.

3. ALGORITHM

This section gives the detail explanation of AVF, WAVF, WDOD, and WADOD methods. The explanation covers the construction of each method and the relationship among them. The explanation is presented in an algorithmic form. A practical example is presented at the end of this section.

3.1. Weighted Attribute Value Frequency (WAVF)

AVF method is one of the simple and efficient unsupervised outlier detection methods for categorical data. AVF method calculates the frequency of each value in each attribute. AVF score of a data point is the average frequency of each attribute of the data point. k outliers are the dataset which have the least k AVF scores [5].

Assume the dataset contain n data points, x_i, $i = 1,\ldots, n$. If each data point has m attributes, $x_i=[x_{i1},\ldots,x_{il}, \ldots x_{im}]$, where x_{il} is the value of the l-th attribute of x_i, below is the AVF score.

$$AVF\ score\ (x_i) = \frac{1}{m}\sum_{l=1}^{m} f(x_{il})$$

where $f(x_{il})$ is the number of times the l-th attribute value of x_i appears in the dataset. k outliers are the dataset which have the least k AVF scores [5].

Algorithm

Input : Dataset – D (n points, m attributes), k target number of outlier
Output : k detected outliers

1. Read dataset D
2. Label all data points as non outliers
3. For each point x_i, i = 1 to n do
   For each attribute l, l = 1 to m do
   calculate frequency $f(x_{il})$ of attribute value x_{il}
4. For each point x_i, i = 1 to n do

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

For each attribute *l*, *l* = 1 to *m* do
AVF score (x_i) += $f(x_{il})$
AVF score (x_i) /=*m*

5. Return *k* outliers with minimum (AVF score)

WAVF improves the performance of AVF method by considering the sparseness of each value attribute. The spareness level is used as the weighting function to the attribute value frequency. The sparseness level of categorical data can be determined by using the statistical function such as standard deviation, variation ratio and range [15].

Instead of using attribute value frequency, WAVF uses attribute value probability. Attribute value probability shows the probability of each value attribute.

The following WAVF algorithm uses attribute value probability range as the weighting function.

Algorithm

Input : Dataset – *D* (*n* points, *m* attributes),
k target number of outlier
Output : *k* detected outliers

1. Read dataset *D*
2. Label all data points as non outliers
3. For each point x_i, *i* = 1 to *n* do
   For each attribute *l*, *l* = 1 to *m* do
   calculate frequency $f(x_{il})$ of attribute value x_{il}
   $p(x_{il}) = f(x_{il})/n$
4. For each attribute a_i, *i* = *1* to *m* do
   $R_i = max(a_i) - min(a_i)$
5. For each point x_i, *i* = 1 to *n* do
   For each attribute *l*, *l* = 1 to *m* do
   WAVF score (x_i) += $p(x_{il})*R_l$
6. Return *k* outliers with minimum (WAVF score)

3.2. Weighted Attribute Density-based Outlier Detection (WADOD)

WDOD uses attribute value frequency as the starting point. Then, WDOD takes into account the degree of infrequent-ness which is measured by using average density. The degree of infrequent-ness makes the attribute give the different effect to the outlier score [8].

Definition

Let DT = (U, A, V, f) be categorical data. For any object $x \in U$, the wieghted density of x in U with respect to A is defined as

$$WDens(x) = \sum_{a \in A} ADens_a(x).W(\{a\})$$

where
$W(\{a\})$ is a weighting function with respect to attribute $a \in A$, given by

$$W(\{a\}) = \frac{1 - E(\{a\})}{\sum_{l \in A}(1 - E(\{l\})}$$

$ADens_a(x)$ is the density of object x in U with respect to the attribute a , given by

$$ADens_a(x) = \frac{|[x]_{\{a\}}|}{|U|}$$

The object *x* in *U* is called a weighted density-based outlier if *WDens(x)* is less then a given threshold value.

Algorithm

Input : Dataset – *D* (*n* points, *m* attributes),
t threshold value
Output : *k* detected outliers

1. Read dataset *D*
2. Label all data points as non outliers
3. For each point x_i, *i* = 1 to *n* do
   For each attribute a_l, *l* = 1 to *m* do
   calculate frequency $f(x_{il})$ of attribute value x_{il}
   $f(x_{il}) = f(x_{il})/n$
4. *Denom* = 0
5. For each attribut a_i, *i* = 1 to *n* do
   P_i=0
   For each value v_{ij}, *j* = 1 to *number of value* of attribute a_i do
   P_i += $f(v_{ij})^2$
   E_i = 1- P_i
   Denom += E_i
6. For each attribute a_i, *i* = 1 to *m* do
   WF_i = (1-E_i) / *Denom*
7. For each point x_i, *i* = 1 to *n* do
   For each attribute a_l, *l* = 1 to *m* do
   $f(x_{il})$*= WF_i
8. For each point x_i, *i* = 1 to *n* do
   $WDens_i$ = 0
   For each atttribut a_l, *l* = 1 to *m* do
   $WDens_i$ += $f(x_{il})$
9. Return points with *WDens* < *t* as outliers

WADOD improves the performance of WDOD by squaring the attribute value frequency. This mechanism widens the range of each attribute value frequency.

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Algorithm

Input : Dataset – D (n points, m attributes),
t threshold value
Output : k detected outliers

1. Read dataset D
2. Label all data points as non outliers
3. For each point x_i, i = 1 to n do
   For each attribute a_l, l = 1 to m do
   calculate frequency $f(x_{il})$ of attribute value x_{il}
   $f(x_{il}) = f(x_{il})/n$
   $f(x_{il}) = f(x_{il})^2$
4. $Denom = 0$
5. For each attribut a_i, i = 1 to n do
   P_i=0
   For each value v_{ij}, j = 1 to *number of value* of attribute a_i do
   $P_i \mathrel{+}= f(v_{ij})^2$
   $E_i = 1 - P_i$
   $Denom \mathrel{+}= E_i$
6. For each attribute a_i, i = 1 to m do
   $WF_i = (1-E_i) / Denom$
7. For each point x_i, i = 1 to n do
   For each attribute a_l, l = 1 to m do
   $f(x_{il}) \mathrel{*}= WF_i$
8. For each point x_i, i = 1 to n do
   $WADOD_i = 0$
   For each atttribut a_l, l = 1 to m do
   $WADOD_i \mathrel{+}= f(x_{il})$
9. Return points with $WDens < t$ as outliers

3.3. Example

The following example shows practical calculation of AVF score, WAVF score, WADOD score. This example is taken from [8]. The calculation of WDOD score, which the original names is WDens score, is presented in [8]. Besides presenting the mechanism, the effect of applying value range to the AVF score and the effect of squaring attribute value frequency to the WDOD score are presented in detail.

Consider categorical dataset in Table 1. The data set has 6 data points: $\{x_1, x_2, x_3, x_4, x_5, x_6\}$. Each data point has 3 attributes: $\{a,b,c\}$. There are three values for attribute a namely {A, B, C}. The frequency of each value is: {2,1,3}. Attribute b has four values namely {D, E, F, G} which have frequency {2, 1, 1, 2}. Attribute c has two values namely {M, N} which have frequency {3, 3}. Table 2 shows the attribute value frequency of this data. Table 3 shows the attribute value probability.

Table 1. Categorical Dataset

U/A	*a*	*b*	*c*
x_1	A	E	M
x_2	A	D	N
x_3	B	G	M
x_4	C	D	N
x_5	C	G	M
x_6	C	F	N

Table 2. Attribute Value Frequency

U/A	*a*	*b*	*c*
x_1	2	1	3
x_2	2	2	3
x_3	1	2	3
x_4	3	2	3
x_5	3	2	3
x_6	3	1	3

Table 3. Attribute Value Probability

U/A	*a*	*b*	*c*
x_1	2/6	1/6	3/6
x_2	2/6	2/6	3/6
x_3	1/6	2/6	3/6
x_4	3/6	2/6	3/6
x_5	3/6	2/6	3/6
x_6	3/6	1/6	3/6

The AVF scores are calculated directly from the entries of Table 2 as follows:

$$AVF(x_1) = \left(\frac{2+1+3}{3}\right) = \frac{6}{3} = 2$$

$$AVF(x_2) = \left(\frac{2+2+3}{3}\right) = \frac{7}{3} = 2.3333$$

$$AVF(x_3) = \left(\frac{1+2+3}{3}\right) = \frac{6}{3} = 2$$

$$AVF(x_4) = \left(\frac{3+2+3}{3}\right) = \frac{8}{3} = 2.6666$$

$$AVF(x_5) = \left(\frac{3+2+3}{3}\right) = \frac{8}{3} = 2.6666$$

$$AVF(x_6) = \left(\frac{3+1+3}{3}\right) = \frac{7}{3} = 2.3333$$

The value range of attribute a, b, and c are :

$$Range(a) = \frac{3}{6} - \frac{1}{6} = \frac{2}{6}$$

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Range(b) = \frac{2}{6} - \frac{1}{6} = \frac{1}{6}$$

$$Range(c) = \frac{3}{6} - \frac{3}{6} = 0$$

Following are the WAVF scores of each point.

$$WAVF(x_1) = (\frac{2}{6} * \frac{2}{6} + \frac{1}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.1389$$
$$WAVF(x_2) = (\frac{2}{6} * \frac{2}{6} + \frac{2}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.1667$$
$$WAVF(x_3) = (\frac{1}{6} * \frac{2}{6} + \frac{2}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.1111$$
$$WAVF(x_4) = (\frac{3}{6} * \frac{2}{6} + \frac{2}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.2222$$
$$WAVF(x_5) = (\frac{3}{6} * \frac{2}{6} + \frac{2}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.2222$$
$$WAVF(x_6) = (\frac{3}{6} * \frac{2}{6} + \frac{1}{6} * \frac{1}{6} + 0 * \frac{3}{6}) = 0.1944$$

The following explanation shows the effect of squaring attribute value probability in Table 3 to the WADOD score. Table 4 is the result of squaring the attribute value probability. The remaining processes are carried out based on Zhao et. Al (2014). The value of *E({a}), E({B}),* and *E({c})* are $\frac{406}{1296}$, $\frac{206}{1296}$, and $\frac{486}{1296}$ respectively. The weighted-density *W({a})*, *W({b})*, and *W({c})* are $\frac{89}{279}$, $\frac{109}{279}$, and $\frac{81}{279}$ respectively. The final AVF score, WAVF score, WDOD score and WADOD score of datasets in Table 1 is shown in Table 5.

Table 4. Squared of Attribute Value Probability

U/A	*a*	*b*	*c*
x_1	4/36	1/36	9/36
x_2	4/36	4/36	9/36
x_3	1/36	4/36	9/36
x_4	9/36	4/36	9/36
x_5	9/36	4/36	9/36
x_6	9/36	1/36	9/36

$$E(\{a\}) = \frac{4}{36}\left(1 - \frac{4}{36}\right) + \frac{1}{36}\left(1 - \frac{1}{36}\right) + \frac{9}{36}\left(1 - \frac{9}{36}\right) = \frac{406}{1296}$$

$$E(\{b\}) = \frac{206}{1296}, \qquad E(\{c\}) = \frac{486}{1296}$$

$$W(\{a\}) = \frac{(1 - \frac{406}{1296})}{\left(1 - \frac{406}{1296}\right) + \left(1 - \frac{206}{1296}\right) + \left(1 - \frac{486}{1296}\right)} = \frac{89}{279}$$

$$W(\{b\}) = \frac{109}{279}, \quad W(\{c\}) = \frac{81}{279}$$

$$WADOD(x_1) = \frac{4}{36} * \frac{89}{279} + \frac{1}{36} * \frac{109}{279} + \frac{9}{36} * \frac{81}{279} = 0.1189$$
$$WADOD(x_2) = 0.1514$$
$$WADOD(x_3) = 0.1249, WADOD(x_4) = 0.1957$$
$$WADOD(x_5) = 0.1957, WADOD(x_6) = 0.1632$$

Table 5. AVF, WAVF, WDOD, and WADOD score

U/A	*AVF score*	*WAVF score*	*WDOD score*	*WADOD score*
x_1	**2**	**0.1389**	**0.3651**	**0.1189**
x_2	2.3333	0.1667	0.4048	0.1514
x_3	**2**	**0.1111**	**0.3492**	**0.1249**
x_4	2.6666	0.2222	0.4603	0.1957
x_5	2.6666	0.2222	0.4603	0.1957
x_6	2.3333	0.1944	0.4206	0.1632

4. EXPERIMENTS

4.1. Experimental Setup

The experiment is done by using Intel Core i5 with 4 GB RAM. The algorithms are implemented in R. The experiment uses three real datasets from UCI Machine Learning repository, namely *Mushroom*, *Nursery*, and *Adult*. The *Mushroom* dataset contains 8124 instances with 23 attributes. The *Nursery* dataset contains 12960 instances with 9 attributes. The *Adult* dataset contains 32561 instances with 15 attributes.

The *Mushroom* dataset is divided into two groups, edible (4208 instances) and poisonous (3916 instances). The edible instances are assumed as the normal data. The poisonous instances are assumed as the outliers. *Nursery* dataset is devided into three groups: usual (4320 instances), pretentious (4320 instances), and great pretentious (4320 instances). The usual instances are assumed as the normal data. The pretentious instances are assumed as the outlier.

The first process in *Adult* dataset is omitting the non- categorical data. Then, the *Adult* dataset is divided into two groups: the persons who have

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

income more than 50K (7841 instances) and the persons who have income less than or equal to 50K (24720 instances). The persons who have income less than 50K are assumed as the normal data, on the other hand, the persons who have more than 50K are assumed as outlier data.

The experimental data setup is carried out by partitioning the outlier data into 1, 2, 4, 8, 16, and 32 partitions. Then, each partition and the normal data are mixed up together, forming the experimental datasets. The performance of the outlier detection algorithm is measured by using the average of detected outlier in each experimental datasets. Table 6, Table 7, and Table 8 show the detected outlier on each experimental datasets by using AVF method, WAVF method, WDOD method, and WADOD method from *Adults* datasets, *Mushroom* datasets, and *Nursery* datasets.

4.2. Result and Discussions

Four outlier detection methods, namely AVF, WAVF, WDOD, and WADOD has been discussed. WAVF is AVF method which uses attribute value probability range as the weighting function. WADOD is WDOD method which uses polinomial orde 2 as the weighting function.

By using categorial data in Table 1, attribute value probability range makes different contribution to the outlier score for attribute *a* and *b* in x_2, though they have the same value of attribute frequency. The different contributions are also given by attribute *a* of x_3 comparing to attribute *b* of x_1 and x_6.

Table 5 shows the comparasion of AVF scores, WAVF scores, WDOD sccores, and WADOD scores completely. Point x_1 and x_3 have AVF score 2 but the WAVF scores are 0.1389 and 0.1111 respectively. The WDOD scores are 0.3651 and 0.3492. The WADOD scores are 0.1189 and 0.1249. Point x_2 and point x_6 have AVF score 2.3333 but the WAVF scores are 0.1667 and 0.1944. The WDOD scores are 0.4048 and 0.4206. The WADOD scores are 0.1514 and 0.1632.

Table 6, Table 7, and Table 8 show the average outlier detected on each partition from *Adult* datasets, *Mushroom* datasets, and *Nursery* datasets respectively. The performance comparation of AVF, WAVF, WDOD, and WADOD on *Adult* datasets are shown in Figure 1.

For the *Mushrooom* datasets, Figure 2 and Figure 3 show the performance comparation of AVF and WAVF, and the performance comparation of WDOD and WADOD respectively. For the *Nursery* datasets, Figure 4 and Figure 5 show the performance comparation of AVF and WAVF, and the performance comparation of WDOD and WADOD repectively.

All the performance comparations show significant performance improvement of AVF and WDOD outlier detection methods by WAVF and WADOD, except WADOD on *Nursery* datasets.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, new outlier detection methods, called WAVF and WADOD have been proposed by applying a weighting function into former outlier detection methods AVF and WDOD. From the experiments, WAVF and WADOD show a significant performance improvement on the outlier detection. WAVF and WADOD can detect more outlier data than AVF and WDOD. Both implementation of WAVF and WADOD on *Adult* dataset give the most significant improvement.

For the future work, the weighting function can be applied to improve the performance of outlier detection method on mixed type dataset. It is possible also to improve the performance WAVF and WADOD by implementing it in a parallel algorithm.

REFERENCES

[1] Grubbs, F.E., 1969, Procedures for Detecting Outlying Observations in Samples, *Technometrics, Volume 11, Issue 1*.

[2] Hawkins, D. M., 1980, *Identification of Outliers*, Chapman and Hall

[3] Phyle, D., 1999, *Data preparation for Data Mining*, Morgan Kaufmann

[4] Chandola, V., Banerjee, A., and Kumar, V., 2009, Anomaly Detection: A Survey, ACM Computing Surveys, Vol 41, No 3., Article 15.

[5] Koufakou, A., Ortiz, E.G., Georgiopoulos, M., Anagnostopoulos, G.C., Reynolds, K.M., 2007, A Scalable and Efficient Outlier Strategy for Categorical Data, *19th IEEE International Conference on Tools with Artificial Intelligence*.

[6] Arif, M., and Basalamah, S., 2012, Similarity-Dissimilarity Plot for High Dimensional Data of Different Attribute Types in Biomendical Datasets, *International Journal of Innovative Compting, Information, and*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Control, Volume 8, Number 2, February 2012, pp. 1275-1297

[7] Shih, M.Y., Jheng, J.W., and lai, L.F., 2010, A Two-Step Method for Clustering Mixed Categorical and Numeric Data, *Tamkang Journal of Science and Engineeering,* Vol. 13, No. 1, pp. 11- 19.

[8] Zhao, X., Liang, J., and Cao, F., 2014, A Simple and Effective Outlier Detection Algorithm for Categorical Data, *Int. J. Mach. Learn & Cyber.* (2014) 5: pp. 469-477.

[9] Chandola, V., Boriah, S., and Kumar, V., 2008, Understanding Categorical Similarity Measures for Outlier Detection, Technical Report, Department of Computer Science and Engineering University of Minnesota, TR 08-008.

[10] Koufakou, A., Secretan, J., Reeder, J., Cardona, K., Georgiopoulos, M., 2008, Fast Parallel Outlier Detection for Categorical Datasets using MapReduce, *IEEE International Joint Conference on Neural Networks.*

[11] Qamar, U., 2013, Automated Entropy Value Frequency (AEVF) Algorithm for Outlier Detection in Categorical Data, *Recent Advances in Knowledge Engineering and Systems Science*.

[12] Reddy, D.L.S., Babu, B.R., and Govardhan, A., 2013, Outlier Analysis of Categorical data using NAVF, *Informatica Economica,* Vol 17, no. 1/ 2013, pp. 5-13

[13] Tan, S.C., Yip, S.H., and Rahman, A., 2013, One Pass Outlier Detection for Streaming Categorical Data, *The 3rd International Workshop on Intelligent Data Analysis and Management*.

[14] Reddy, D.L.S. and Babu, B.R., 2013, Outlier Analysis of Categorical data using FuzzyAVF, *International Conference on Circuits, Power and Computing Technologies, (ICCPCT 2013)*.

[15] Blalock, H.M., 1981, *Social Statistics*, McGraw-Hill.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6. Outlier Detected from Adult Datasets

partition	Outlier	Average outlier detected			
		AVF	WAVF	WDOD	WADOD
1	7841	2093.0	1984.0	2185.0	2397.0
2	3920	1093.5	1353.0	1307.0	1500.0
4	1960	370.3	483.8	621.3	878.0
8	980	147.9	175.3	198.3	205.0
16	490	56.8	77.0	87.6	93.7
32	245	18.9	24.9	29.7	38.6

Table 7. Outlier Detected from Mushroom Datasets

partition	Outlier	Average outlier detected			
		AVF	WAVF	WDOD	WADOD
1	3916	2128.0	2254.0	2230.0	2273.0
2	1958	1007.0	1027.5	998.0	1014.5
4	979	611.3	628.5	601.8	600.8
8	489	302.6	303.8	287.1	287.9
16	244	158.4	169.7	168.3	174.3
32	122	82.1	90.0	89.0	92.1

Table 8. Outlier Detected from Nursery Datasets

partition	Outlier	Average outlier detected			
		AVF	WAVF	WDOD	WADOD
1	4320	2381.0	2381.0	2381.0	2381.0
2	2160	1583.0	2159.0	2160.0	2160.0
4	1080	1069.0	1080.0	1080.0	1080.0
8	540	540.0	540.0	540.0	540.0
16	270	270.0	270.0	270.0	270.0
32	135	135.0	135.0	135.0	135.0

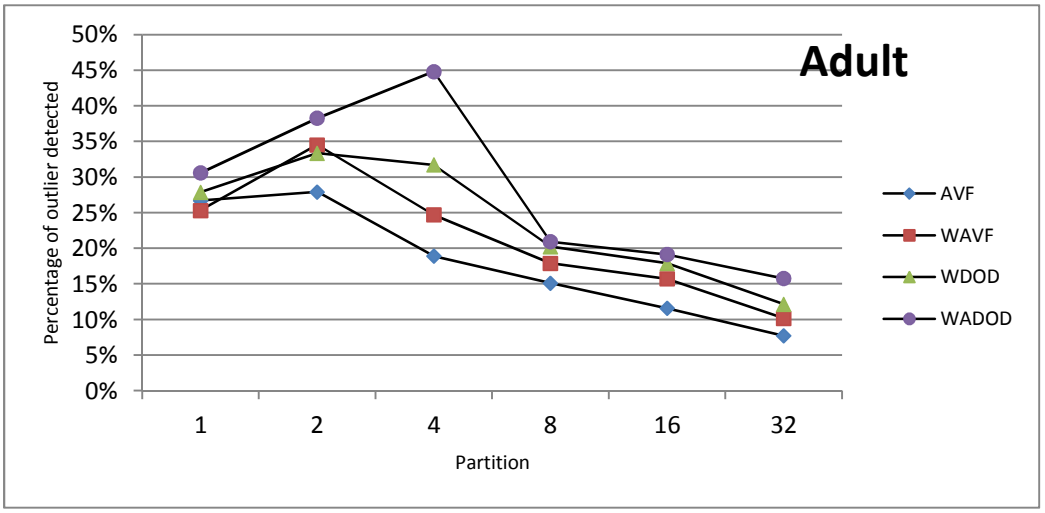

Figure 1. Performance Comparasion of AVF, WAVF, WDOD and WADOD on Adult Datasets

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

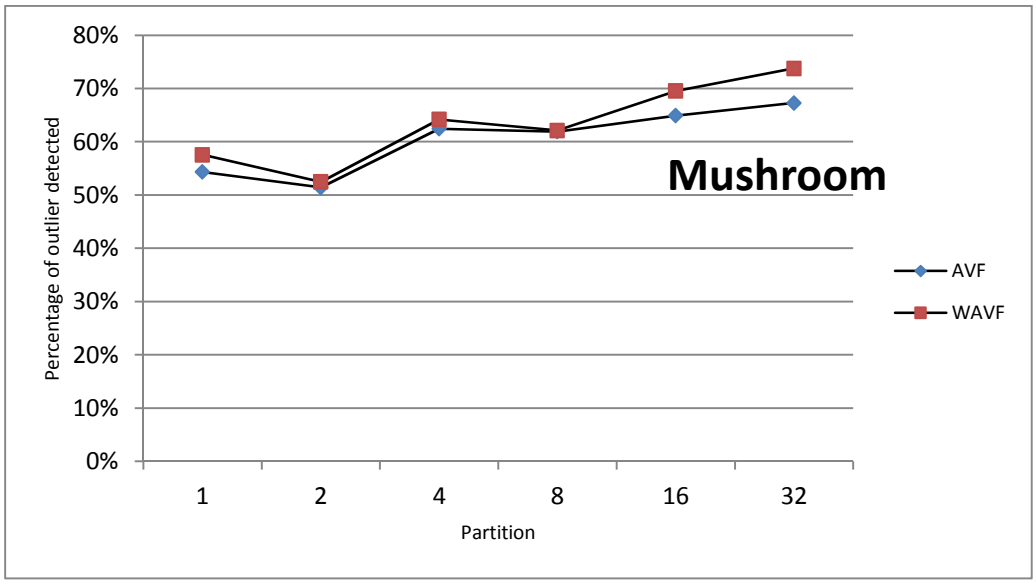

Figure 2. Performance Comparasion of AVF and WAVF, on Mushroom Datasets

Figure 3. Performance Comparasion of WDOD and WADOD on Mushroom Datasets

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

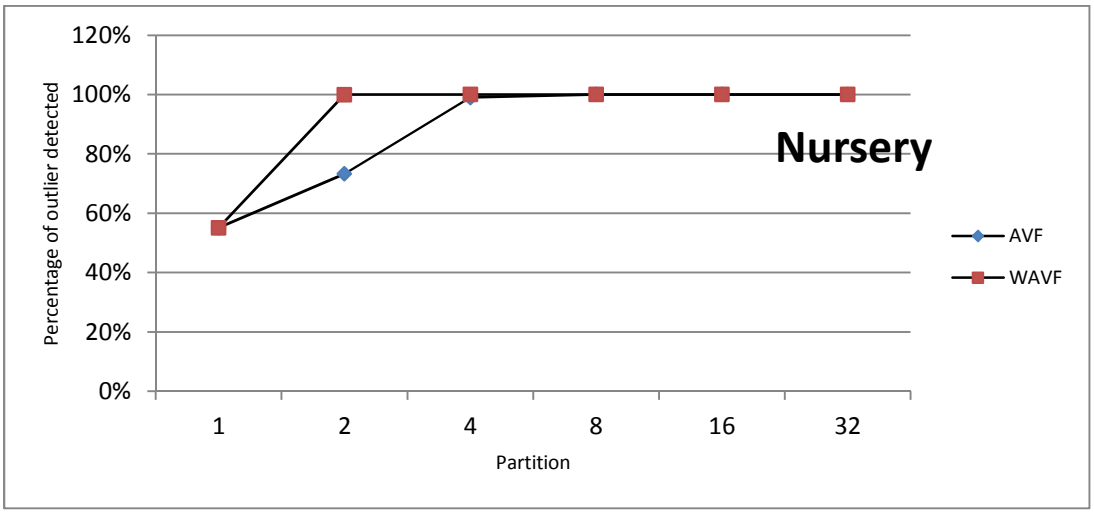

Figure 4. Performance Comparasion of AVF and WAVF on Nursery Datasets

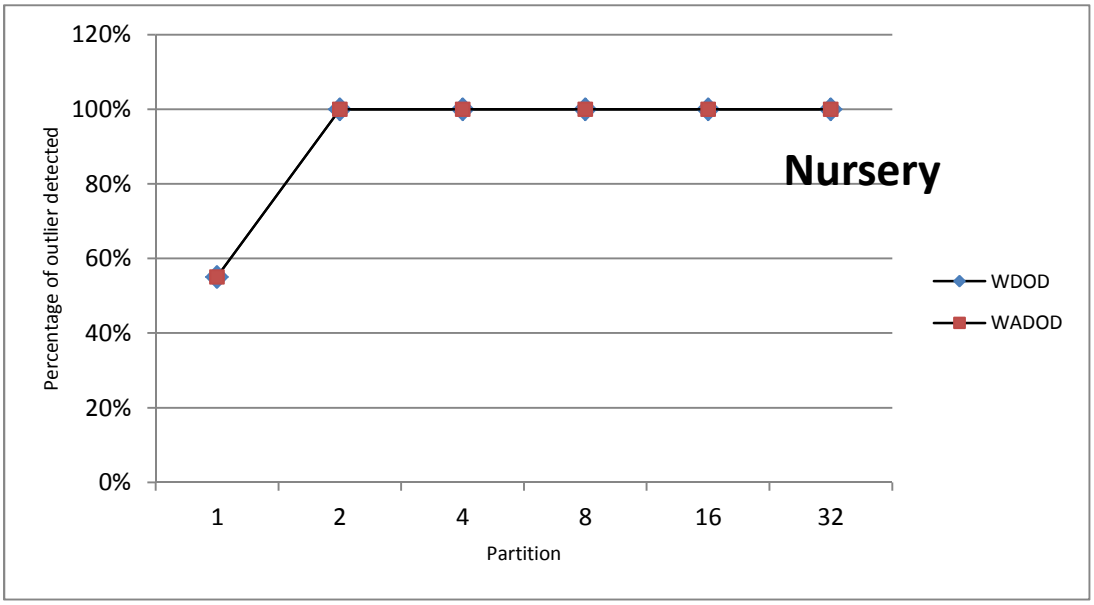

Figure 5. Performance Comparasion of WDOD and WADOD on Nursery Datasets