

## Binary Priority Outlier Classifier Based Outlier Elimination

Deoras Tejas Tushar<sup>a</sup>, Senthilnathan P<sup>b\*</sup>, K. Deeba<sup>c</sup>, N. Venkata Vinod Kumar<sup>d</sup>

<sup>a,b,c</sup>School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

<sup>d</sup>Department of C.S.E, Annamacharya Institute of Technology and Sciences, Tirupati, India

<sup>b\*</sup>senthilnathan.p@vit.ac.in

**Article History:** Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

**Abstract:** Outliers are records that deviate from normal behavioral pattern. This causes a serious issue when it comes to analysing data. In the recent years there has been great research to identify these outliers. Identifying them not only helps improve analysis of data but also provides many applications. The paper presents a way of indenting these outliers based on priority assigned to the attributes. The priorities are then added for each record in the dataset and the pattern is analysed. A concept based on interquartile range is used to eliminate the outliers. Hence the classifier divides the dataset into two classes: outliers and normal data.

**Keywords:** Outliers, Classifier, Interquartile range

### 1. Introduction

Outlier detection has been used since a long time to detect anomalous behavior. Outliers are caused by faults in machines, frauds, human error or simply a phenomenon of natural deviations. Identifying them provide extremely valuable information and early identification of them can help prevent catastrophic consequences. Victoria Hodge et al. discuss various algorithm for outlier detection and compare them by analysing their advantages and disadvantages [1]. Robert.L.Lipnik et al. discuss how molecular descriptors and molecular toxicity can help identify the outlier behaviour as well as provide information about the predictive capability of such models. The paper uses QSAR baseline prediction and compares it with toxicity levels for identifying outliers and impress the classification process by removing the same[2]. Yang Zing et al. discuss how the identification of outliers in wireless sensor networks can provide valuable information such as noise, errors and malicious attack affecting the network. Traditional outlier methods fail when used on wireless networks due to various requirements and limitations specific to networks. The paper provides an algorithm based on taxonomy and comparative table to select a technique from the available wireless network outlier detectors based on data type, outlier type, outlier identity and its degree[3]. Jorm Laurikkala et al. discuss the identification of outliers with the help of box plot in the field of medicine. They plotted the distance on box plot using Mahalanobis distances to identify multivariate outliers and univariate outliers were detected directly using a box plot. The identification of outliers not only helped increase the predictive ability of the classification but also the most experts in the field actually recognized the records to be outliers in their area[4]. Sofie Verbaeten at al. uses outlier detection as an application in noisy training examples where certain records are mislabelled. They use outliers methods as a pre-processing task and then proceed with the classification process. They use a number of filtration techniques like cross validation, boosting and bagging. They evaluate these techniques in an Inductive Logic programming setting and use decision tree to construct these ensembles[5]. Outliers provide valuable information about the data and should not be ignored. Identifying them not just provides various applications but it also provides industries the ability to get a clean dataset and identify the normal patterns in the data.

### 2. Methodology

The algorithm uses posterior likelihood in Bayesian statistics as a measure for assigning priority to each attribute.

Posterior likelihood – In Bayesian Statistics, the posterior likelihood of a random event or associate degree uncertain proposition is that the chance that's assigned when the relevant proof or background is taken under consideration.

Let us have a prior belief that the probability distribution function is  $P(\Theta)$  and observations  $X$  with the likelihood then the Posterior likelihood is:

$$P\left(\frac{\theta}{X}\right) = \left(P\left(\frac{X}{\theta}\right) * P(\theta)\right) / P(X)$$

The outliers are eliminated by interquartile range. IQR stands for ‘interquartile range’. It is used in statistics to help analyse set of numbers. IQR is preferred over range as it is better at identifying outliers.

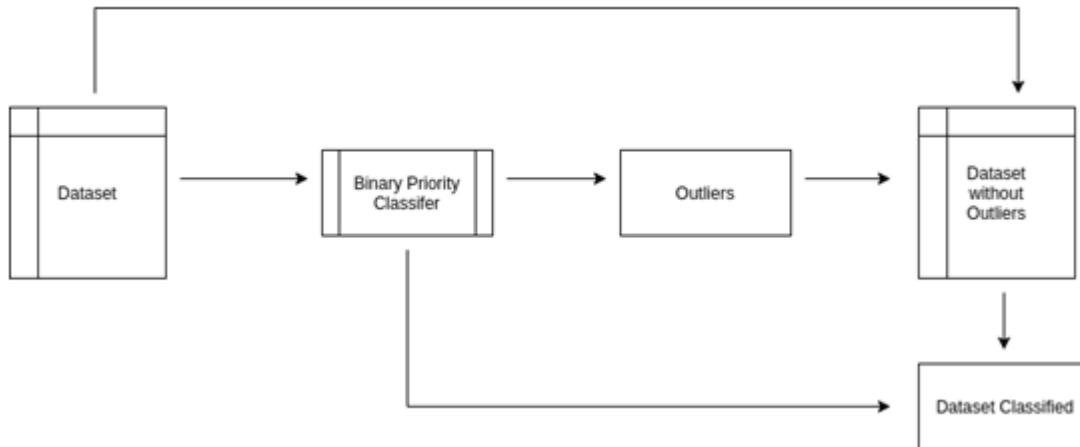


Fig 1. Workflow

**Work Flow:**

A dataset is first imported for outliers detection.

The Binary Outlier Classifier identifies the Outliers.

A dataset without the outliers is generated.

$$D(0 \dots N) \rightarrow A(0 \dots K) \quad (1)$$

For each attribute in A,

$$P\left(\frac{L}{A}\right) = \left(P\left(\frac{A}{L}\right) * P(\theta)\right) / P(X) \quad (2)$$

$$A(0 \dots K) \rightarrow \textit{Ascending order sort based on } P(L/A)$$

Assign each attribute a priority based on P (L/A) from 0 to K in increments of 1.

For each training data records imported calculate the sum of priorities for the attributes that occur in that record.

$$H(0 \dots M) = \sum_{\epsilon \text{ record}} \textit{Each attribute occuring value} \quad (3)$$

$$H(0 \dots M) \rightarrow \textit{Ascending order sort}$$

Calculate interquartile range for H(0...M)

$$IQR=Q3-Q1 \quad (4)$$

Find farthest entries in H (0...M)

$$\text{If } H(0 \dots M) > 1.5 \times IQR \quad (5)$$

Find corresponding record and remove it from dataset as it an outlier.

$$M=M-O \quad (6)$$

Symbols:

D (0...N) - Dataset with N records A (0...K) – Record with K attributes

H (0...M) – Sum of priorities for each record in training set of point M L: Label in the dataset

The above algorithm is only used for identifying outliers for a single label. The attributes for other labels can be calculated in similar way.

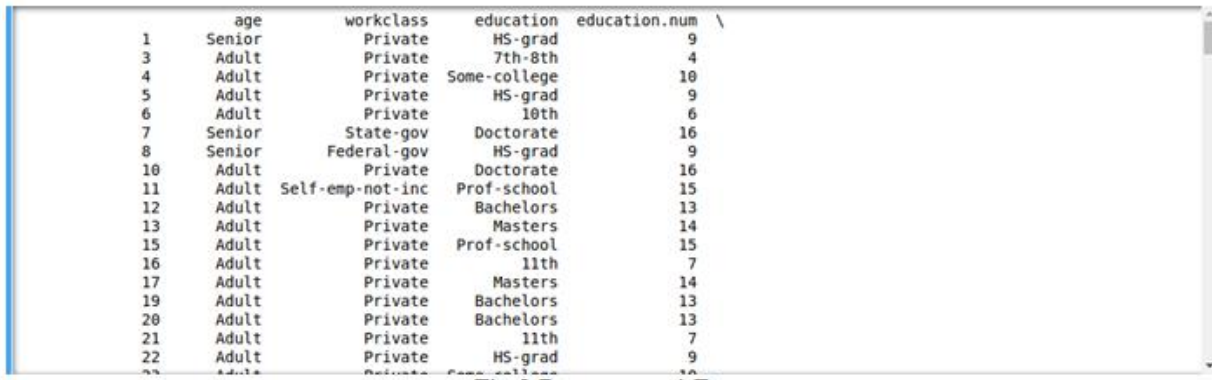
### 3. Data Set & Tool used

Adult Census Income dataset is downloaded from the UCI Machine Learning Repository. Total number of instances are 48842, number of attributes are 14 and it is a multivariate data set. Attributes are Categorical, Integer. Tool used is Jupyter Notebook. The Jupyter Notebook is net application that enables you to make and share documents that contain live code, equations, visualizations and narrative text. Uses include: information cleansing and transformation, numerical simulation, applied mathematics modeling, information visual image, machine learning, and a far lot. The Notebook can be used to code in many languages, including Python, R, Julia, and Scala.

### 4. Experimental Result and Discussion

#### Data Pre-processing

One of the main shortcomings of the given algorithm is that it only works for categorical data. The reason for this is because the algorithm needs to calculate posterior probability for each distinct attribute. For continuous attributes the no of distinct values may be extremely large. Hence it is important to convert this continuous data into categorical data. The attributes age, capital gain, capital losses are converted into categorical data. Missing values are also removed.



	age	workclass	education	education.num	\
1	Senior	Private	HS-grad	9	
3	Adult	Private	7th-8th	4	
4	Adult	Private	Some-college	10	
5	Adult	Private	HS-grad	9	
6	Adult	Private	10th	6	
7	Senior	State-gov	Doctorate	16	
8	Senior	Federal-gov	HS-grad	9	
10	Adult	Private	Doctorate	16	
11	Adult	Self-emp-not-inc	Prof-school	15	
12	Adult	Private	Bachelors	13	
13	Adult	Private	Masters	14	
15	Adult	Private	Prof-school	15	
16	Adult	Private	11th	7	
17	Adult	Private	Masters	14	
19	Adult	Private	Bachelors	13	
20	Adult	Private	Bachelors	13	
21	Adult	Private	11th	7	
22	Adult	Private	HS-grad	9	

Fig 2. Preprocessed Data

#### Implementing the algorithm

The pandas library in python is used for implementation. Pandas provide statistical and data mining tools for coding in python. The implementation first starts with calculating the conditional probability of each attribute and placing them in ascending order. After that based on assigned probability a weight is calculated for each similar label record. On implementation it is quite obvious that records with similar labels will have weights close by. Hence a interquartile range value can be used to identify weights which are far away.

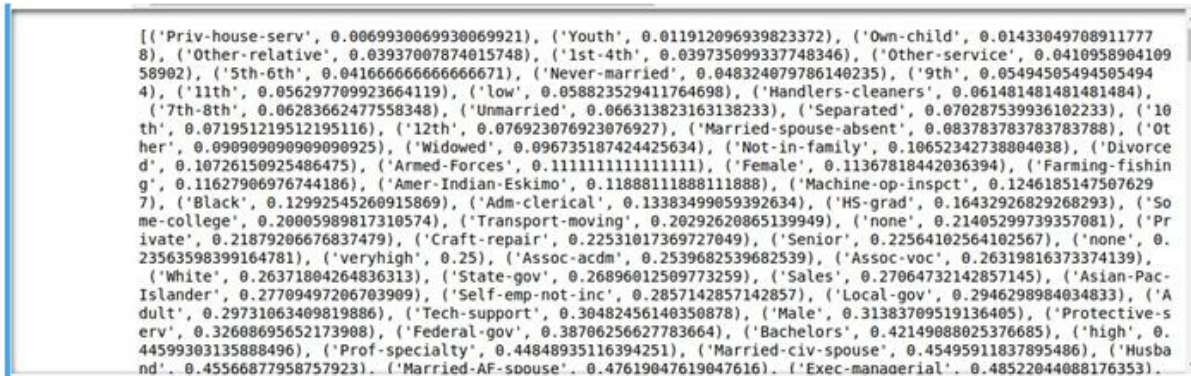


Fig 3.Attributes with conditional probability in ascending order



Fig 4.Weights of records with label(>50k) with corresponding record number.

The IQR (Inter quartile range) = `48

No of records with label(>50k)	Outliers identified	Outliers percentage
7508	246	3.27%

Table 1.No of outliers identified

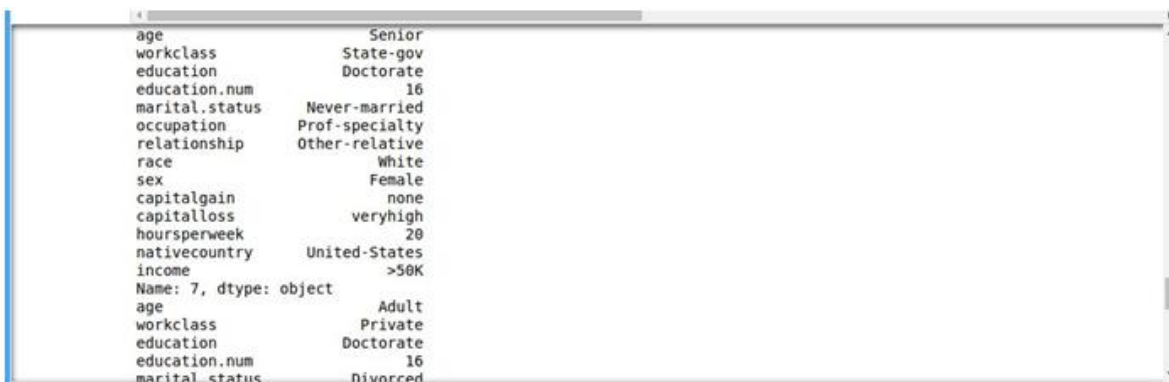


Fig 5.Outlier in Dataset

5. Conclusion

This paper has presented an algorithm for identifying outliers. As we saw that the algorithm successfully classified 3.27% of the dataset as outlier. While the number may seem small the records provide extremely valuable information. By using the posterior probability to assign weight for each record the algorithm created clusters and eliminate those as outliers which fall outside them. Hence the algorithm is successful in identifying outliers for any dataset provided that dataset has categorical data.

## References

- Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2),85-126.
- Lipnick, R. L. (1991). Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Science of the total environment*, 109,131-153.
- Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials*, 12(2),159-170.
- Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., & Kavsek, B. (2000, August). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology* (Vol. 1, pp.20-24).
- Verbaeten, S., & Van Assche, A. (2003, June). Ensemble methods for noise elimination in classification problems. In *International Workshop on Multiple Classifier Systems* (pp. 317-325). Springer, Berlin, Heidelberg.
- Gopichand, G., & Saravanaguru, R. (2016). A Generic Review on Effective Intrusion Detection in Ad hoc Networks. *International Journal of Electrical and Computer Engineering*, 6(4), 1779.
- G. Gopichand, R.A.K. Saravanaguru, K. Ramesh Babu, Fully Secured Intrusion Detection System For Sensing Attacks in MANET, *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 4 Special Issue, pp. 810-816, 2018
- Gopichand Ginnela and Ramaiah Kannan Saravanaguru, “Collaborative Packet Dropping Intrusion Detection in MANETs”, *Recent Advances in Computer Science and Communications* (2020) 13(6).
- Gopichand G., Sankeerth K.S., Parlapalli A, Evaluation of recommendation systems using trust aware metrics, *International Journal of Recent Technology and Engineering*, Volume-7, Issue- 6S4, April 2019
- Gopichand, G., Jain, K., Dev, S.K, Research on e-healthcare security evaluation in cloud-based system, *International Journal of Recent Technology and Engineering*, Volume-8, Issue-2S11, September 2019
- Gopichand G, Vishal Lella, Sai Manikanta Avula, Enhancing Performance of Map Reduce Workflow through H2HADOOP: CJBT, *International Journal of Recent Technology and Engineering*, Volume-7, Issue-6S4, April 2019
- Gopichand G, Sailaja G, N. VenkataVinod Kumar, T. Samatha, Digital Signature Verification Using Artificial Neural Networks, *International Journal of Recent Technology and Engineering*, Volume-7 Issue-5S2, January 2019
- Gopichand G, Ra.K.Saravanaguru, .K.Ramesh Babu, Usage of AODV and AOMDV Protocols in Perceiving Black hole Attacks in a MANET, *International Journal of Pharmacy & Technology*, Volume 8, Issue 4, December 2016
- Gopichand G, Ra.K.Saravanaguru, K.Ramesh Babu, , MITIGATING DDOS ATTACKS THROUGH AODV PROTOCOL IN A MANET USING NS3 SIMULATOR *International Journal of Pharmacy & Technology*, Volume 8, Issue 4, December 2016
- Gopichand, G., Vamsi, K.S.V., Subhash Reddy, Y.S., Chand, K.S.P., Saiteja, G, A hybrid scheme in cloud computing for secure sharing of data in the cloud, *International Journal of Recent Technology and Engineering*, Volume-8, Issue-2S11, September 2019
- Sreekant, A., Senthilnathan, P., Gopichand, G., Rajapandy, M., Kannan, N. Necessity of machine learning and data visualization principles in marketing investment management , *International Journal of Innovative Technology and Exploring Engineering*, Volume-8, Issue- 6S4, April 2019
- Mehta M., Rajesh Mamilla, Sunithavenugopal, Gopichand G, Growth and development of start- ups in India - A study with respect to mechanical and production engineering, *International Journal of Mechanical and Production Engineering Research and Development*, Volume : 8-2, April 2019
- H R Swathi, Shah Sohini, Surbhi, Gopichand G, Image compression using singular value decomposition, *IOP Conference Series: Materials Science and Engineering* 263(4).
- Gopichand, G., Sola, K., Reddy, C. B. S., Rakesh Kumar, M. V., & Vardhan, H. Vocabulary Mismatch Avoidance Techniques.
- Priyadarsini, M. J. P., Rajini, G. K., Naseera, S., Balaji, S., Reddy, P. S. K., & Gopichand, G. (2006). AUTOMATIC OBJECT RECOGNITION BASED ON EUCLIDEAN DISTANCE RESTRICTED AUTO ENCODER.
- Palaniappan, S., Palli, S., Ginnela, G., Ameerjohn, S., & Gopal, S. S. (2018). Enhanced Handwritten Number Detection Using Kernel Discriminant Analysis (KDA). *Journal of Computational and Theoretical Nanoscience*, 15(8), 2539-2543.

Shaw, J., Vincent, P. M., Paliappan, S., Sangaiah, A. K., & Gopichand, G. (2018). Intelligent Phishing Detection System Using Feature Analysis. *Journal of Computational and Theoretical Nanoscience*, 15(8), 2533-2538.