Please select category below:
Normal Paper ☒
Student Paper ☐
Young Engineer Paper ☐

Categorical outlier detection for Health Usage and Monitoring Systems

Leonard Whitehead

Aerospace Division, Defence Science and Technology Group,
506 Lorimer Street, Fishermans Bend, Victoria, 3207, Australia

Abstract

The rise of aircraft health and usage monitoring systems (HUMS) data volume and datatype variety, precipitates the opportunity and imperative to increase t he analyst toolkit. Applying outlier detection methods to HUMS data can inform the user of a series/sequence of observations that are indicative of a failure within the system. This paper explores outlier detection for categorical data using algorithms based on frequency- attribute value frequency (AVF); entropy-automated entropy value frequency (AEVF) and probability-conditional algorithm (CA). Exploring these algorithms has achieved a deeper understanding of the properties involved in outlier detection for categorical data.

Keywords: Anomaly detection, categorical data, frequency, Bayesian probability, entropy.

Introduction

Health and usage monitoring systems (HUMS) data structures are growing in complexity and subsequently the abnormality detection methodologies applied must also transform. Detecting an outlier as an **"**observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [1] is of high interest to platform health monitoring. Detection of outliers in categorical data is difficult compared to numerical data. Unlike numerical data, categorical outlier detection focuses on an abnormal pattern (sequence) or a rare (singular) occurrence of classes within a feature whereas numerical data may focus on the distance between data points. Attribute value frequency (AVF), conditional algorithm (CA) and automated entropy value frequency (AEVF) are categorical outlier detection algorithms that belong to the frequency and Bayesian/conditional hierarchies [2]. Through the exploration and testing of these algorithms, insights of the categorical outlier detection field have been gleaned.

Methodology

In order to test the algorithms, a dummy data set was created using the programing language Python. The dataset has three features - altitude, condition code and health - where each feature has a different amount of classes. Each class has a different probabilistic weighting

attached so that when the data were randomly generated, some classes appear more than others, as shown in Tables 1, 2 and 3.

Table 1 Counts of Altitude classes

Altitude Class	Count
Medium	296
Low	181
High	209
Ground	78
Very Low	73
Very High	99
Too high	64

Table 2 Counts of Health classes

Health Classes	Count
Normal	500
Good	234
Bad	158
Super good	46
Extremely bad	62

Table 3 Counts of condition classes

Condition Classes	Count
Engine-on	315
Radio-on	198
Temperature-norm	188
Temperature-high	107
Temperature-freezing	55
Radio-off	52
Engine-off	44
Temperature-critical	41

Discussion

Each algorithm was randomly chosen as the field of categorical outlier detection was explored. Through research and testing, the time complexity (big O notation) and characteristics of most algorithms were discovered. While the AVF and CA are both reviewed by [2] the AVEF is not; however, its hierarchy-entropy is.

An introductory and summarized view of each algorithm is below along with a scenario. As each algorithms specifications, have been published in their respective papers, these will not be discussed below.

Scenario

As an aircraft is undergoing a mission, it sends the pilot a series of messages which contain its altitude, condition code and health status. The pilot must analyse each message and look out for potential anomalies. To assist with the anomaly detection, the pilot has a series of algorithms which generate additional information which help determine if an anomaly has occurred.

Attribute Value Frequency (AVF)

The AVF algorithm calculates the sum of each class for each feature for every record (row in a dataset) divided by the total amount of features. The k records with the lowest AVF score are then designated as outliers [5].

$$AVF(x) = \frac{1}{m}\sum_{1}^{m} f(x_i) \qquad (1)$$

Where

- m = number of columns
- x_i = The count of the class in cell *i*th column in row x.

Frequency-based methods are fast and scalable, but they ignore any dependency structure among categorical variables and the number of anomalies are set in advance [2].

Table 4 below highlights how the outliers are not dependent on the sequence/structure of categorical variables but instead the individual low frequency of each item. This leads to masking (failing to declare some outliers) in the AVF. For example, the record "High, Temperature-critical, normal" has an AVF score of 250 even though the structure of the record is clearly an outlier.

This problem can be resolved by counting the records instead of the classes in each feature, however then the records that are labelled as outliers will likely have classes that are not outliers.

Table 4 AVF scores, low to high, shortened

altitude	Condition code	health	Row total
Very Low	Temperature-critical	super good	53.33
Too High	Temperature-critical	extremely bad	55.66
Ground	Engine-off	super good	56
Very Low	Temperature-critical	extremely bad	58.66
…	…	…	…
High	Temperature-critical	Normal	250

Automated Entropy Value Frequency (AEVF)

Qamar (2013) describes the AEVF as a two-step process.

1) Calculate the entropy change values and then reorder the dataset so that the entropy change values are in descending order from highest to lowest.
2) Determine the outliers by calculating the entropy difference gap for each record, if the entropy difference gap is equal or greater than the maximum entropy gap then terminate the algorithm. All values above the record on which the algorithm terminated are then considered outliers. The formula for entropy when the features are independent is below:

$$H(x) = -\sum_{x=1}^{n} p(x)\log(p(x)) \qquad (2)$$

Entropy difference gap: The entropy difference between two records.
Maximum entropy gap: A value usually set by the user, in this case it is instead the average value of the entropy change values.
Entropy change value: This is the difference between the main entropy of the data set and the new entropy of the dataset when a record is removed.

Where x is the data point/row

The AEVF has the following weaknesses, which are shared with other entropy-based algorithms [2]:

- Need to identify the number of outliers in advance (setting max entropy gap)
- High time complexity
- Lots of ties due to same minimum entropy being calculated
- Subject to masking and swamping (labelling normal events as outliers).

Unlike other entropy algorithms, the AEVF uses a maximum entropy gap which is less intuitive then setting the amount of outliers to be designated [3]. Table 5 below shows some of the outliers.

Table 5 Outliers according to entropy

altitude	Condition code	health	entropy gap
Too High	Engine-off	normal	0.004
High	Radio-off	super good	0.004
Very Low	Temperature-critical	normal	0.004
Ground	Temperature-norm	bad	0.004

Conditional Algorithm (CA)

The CA calculates a ratio of the joint probability of a data point divided by its marginal probability. Records that have a high ratio value indicate a suspicious coincidence of events co-occurring (Table 6), while, low values signify that the events do not co-occur naturally (Table 7) [4], and the values in between represent normalcy.

Ratio values can be determined to be high or low by taking a percentage of the uppermost and bottommost values or taking the values outside a certain range from the mean. Validation of the suspicious values and events that do not co-occur naturally is helpful in determining the ratio boundaries for upper and lower values.

$$r(a_t, b_t) = \frac{p(a_t, b_t)}{p(a_t) * p(b_t)} \quad (3)$$

Where a_t and b_t represent different events or a group of events. For example, the ratio of it being hot (a_t) with an overcast of cloudy (b_t).

While the CA does not have to set the amount of outliers like the frequency and entropy based methods, it instead has time complexity problems [2].

Table 6 High ratio – suspicious coincidence

altitude	Condition code	health	ratio
Ground	Engine-off	good	13.69
High	Temperature -critical	extremely bad	9.41
Medium	Temperature -critical	extremely bad	9.30
Ground	Engine-off	bad	9.22
Very High	Temperature -critical	extremely bad	7.94
Ground	Engine-off	normal	7.57

Table 7 Low ratio – non natural co-occurring events

altitude	Condition code	health	ratio
Ground	Engine-on	bad	0.257
Low	Temperature -high	bad	0.326
Low	Radio-on	bad	0.353
High	Temperature -freezing	good	0.371
Low	Engine-on	super good	0.381

Another drawback is shown in Table 7 that non-natural co-occurring records are not necessarily outliers. They are records that rarely happen but their classes (events) are quite frequent throughout the data set.

Determining which algorithm to use

When deciding which algorithm to use, the time complexity, type of learning, output and accuracy of the algorithm, are key factors that the user needs to take into consideration. For example, the AVF is faster than the CA due to its low time complexity in comparison to the CA. However, the output of the AVF compared to the CA is not as rich in information, the CA can inform the user of suspicious coincidences or non-natural coincidences, whereas the AVF can only supply a frequency-based score. Overall the AVF is fast but lacks rich information, the CA is slow but generates rich information. In this case the user needs to weigh up which factors they value most to decide which algorithm is best suited to meet their objective.

To help determine which categorical outlier methods may be applicable, a list of simple criteria is proposed below:

- Does the algorithm meet your objectives?
- Does the algorithm match your time complexity needs?
- Does your data meet the required assumptions of the algorithm?
- Does the algorithm discover the “type” of outliers you are searching for? (Dependency structure vs individual or both)
- Do the algorithm parameters suit your needs? (i.e. set amount of outliers)

After using the criteria to create a set of potential methods/algorithms, tests can be done to validate the accuracy, speed, and other key factors of the algorithms, to determine which one is best suited to the user’s needs.

Conclusion

This paper has explored a range of anomaly detection techniques for use with aircraft HUMS categorical data. There are many different categorical outlier detection algorithms each with their associated hierarchy and unique value. Using the criteria mentioned in the discussion, a set of potential algorithms can be determined, whereas further testing can lead to the most suitable algorithm being chosen for a specific application. Implementing categorical outlier detection methods into HUMS data is very useful and has the potential to enhance platform state awareness and by providing information to pilots, operators and maintainers where intervention may be required. Future research will explore the application of these techniques to aircraft flight data to enhance autonomous system state awareness.

References

1. Hawkins,D.M.(1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.
2. Taha, A., & Hadi, A. S. (2019). Anomaly detection methods for categorical data: A review. *ACM Computing Surveys (CSUR), 52*(2), 1-35.
3. Qamar, U, (2013). Automated entropy value frequency (aevf) algorithm for outlier detection in categorical data. *Recent Advances in Knowledge Engineering and Systems Science.*
4. Das, K., & Schneider, J. (2007, August). Detecting anomalous records in categorical datasets. *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 220-229).
5. Koufakou, A., Ortiz, E. G., Georgiopoulos, M., Anagnostopoulos, G. C., & Reynolds, K. M. (2007, October). A scalable and efficient outlier detection strategy for categorical data. *In 19th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI 2007),2007,pp.210-217, doi: 10.1109/ICTAI.2007.125.