Anomaly Detection using Similarity Approach on Airline Data

Utpal Kumar Sikdar*, Krishna Kumar M

IBS Software Pvt Ltd. Trivandrum, Kerala, India

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.15

ABSTRACT

Anomaly detection is to identify abnormal items, events or observations from the majority of the data. We applied similarity approaches to identify the abnormal observations from the Airline Data on chargeable weight. Chargeable weight is what the airline uses to determine the cost of the shipment. It may be either volumetric weight or gross weight, whichever is greater. Similarity approaches are applied to identify the abnormal observations on chargeable weight and evaluated the systems with the airline data. The precision, recall and F-measure values of the best system are 41.12%, 54.91% and 47.02% respectively.

Keywords: Simple Similarity Approach (SSA), Majority Voting Similarity Approach (MVSA), Weighted Voting Similarity Approach (WVSA).

1 Introduction

In data analysis, anomaly detection is the identification of rare observations which raise suspicions by differing significantly from the majority of the data. Typically, the anomalous observation will translate to some kind of problem such as financial fraud detection [6], a structural defect [1], medical problems [2] or errors in a text. There are three broad categories of anomaly detection techniques exist, namely unsupervised, supervised and semi-supervised. Unsupervised anomaly detection is to identify anomaly data from the unlabelled dataset under the assumption that some of the data points are differ from the majority of the data set. In unsupervised approach, a set of rules are applied to identify the anomaly data. In supervised anomaly detection techniques require a data set that has been labelled as "normal" and "abnormal" and involves training a classifier. Semi-supervised anomaly detection techniques construct a model that is combination of supervised and unsupervised approaches.

There are a few work exist on Airline data for anomaly detection. In the paper [5], the authors presented an approach for identifying anomalous flight data records from General Aviation operations. The authors also described a density-based clustering and one-class classification methods together for anomaly detection using energy-based metrics. A cluster analysis technique [4] has been developed to support Flight Operations Quality Assurance (FOQA) by identifying anomalous flights based on onboard-recorded flight data. The method generated a high dimensional data vector on time series data from multiple flight parameters and applied to identify anomalous flights using cluster based techniques. The paper [3] demonstrated an automated data processing approach for finding subtle anomalies in aircraft performance from very large Flight Operations Quality Assurance (FOQA) data sets. Statistical Process Control (SPC) has been applied to identify anomaly on FOQA data.

As per our knowledge, there is hardly any work done on chargeable weight for anomaly detection. The chargeable weight is very important with respect to cargo revenue management. If the chargeable weight is wrongly applied on different agent for different AirWayBill (AWB) shipment Origin and Destination pair (OD-

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

pair), there is a huge impact on revenue. Chargeable weight is decided based on AWB shipment OD-pair for each individual commodity code. In this paper, a similarity based approach was applied on chargeable weight to identify anomaly from the Airline data. We also applied majority and weighted voting approaches for anomaly detection.

The rest of the paper is organized as follows: The anomaly detection techniques are introduced in Section 2. Data and Results are presented and discussed in Section 3, while Section 4 addresses Error Analysis. Lastly Conclusion is addressed in Section 5.

2 Anomaly Detection Approach

The anomaly detection approaches are applied on Airline data. To identify the anomaly, several similarity approaches were tested on label tagging converted data. The label '1' represents the changes between initial and final values of the chargeable weights called as anomaly. The label '0' represents as non-anomaly where there will be no changes between initial and final values of the chargeable weights. Here initial chargeable weight means the assigned chargeable weight when a new booking has been done. Later the initial chargeable weight may be updated (called final chargeable weight) if any suspicions value is found in initial chargeable weight.

2.1 Feature Description

The similarity approaches were developed on airline data using different features. The following features are applied on similarity approaches:

The features are categorized into three types:

- Numeric Features
- Single Valued Categorical Feature
- Multi-valued Categorical Feature.

2.1.1 Numeric Feature

Numeric features are identified from a set of features based on the numerical values. For example, an initial gross volume (GRSVOL) feature is a numeric feature because the feature contains only numeric value for each data point. The initial gross volume (GRSVOL) is directly associated with the chargeable weight. In general more chargeable weight shipment is having more volume. Numeric feature descriptions are mentioned below:

1. Chargeable Weight (CHGWGT): The Airlines define chargeable weight for a particular shipment route called 'AirWaybill (AWB) OD-pair'. The chargeable weight value for a particular AWB OD-pair may change during audit process if any suspicious value is found on initial chargeable weight.
2. Gross Volume (GRSVOL): During the AWB booking, the gross volume is maintained for a particular shipment.
3. Gross Weight (GRSWGT): During the AWB booking process, gross volume is also maintained for a particular shipment.
4. Total Market Charge (TOTMKTCHG): The AWB contains the total market charge for a particular shipment.

2.1.2 Single Valued Categorical Feature

In single valued categorical feature, there is a single value for each data point. For example, there are total seven product codes in 'PRDCOD' feature but each data point is having only one single product code value. The single valued features are listed below:

1. Product Code (PRDCOD): Each of the shipment is associated with product code. Product code identifies the shipment categories for each commodity code.
2. Agent Code (AGTCOD): In general, most of the shipments are booked by the agents for a particular flight. Each agent for a particular shipment route booked a set of product which are almost same product types.
3. Booking Currency Code (CURCOD): Each of the booking is having the currency code. When AWB shipments are booked, the booking currency code may be different based on the location/region/county.

2.1.3 Multi-valued Categorical Feature

There are multiple categorical values for each data point. The multiple categorical values are separated by comma in each multi-valued categorical feature. Each of the multi-valued categorical features is described below. For example, in Special Handling Code (SCCCOD) feature, one of the data point value is 'AVI,BIG,P31'. Here three different categorical values are 'AVI', 'BIG' and 'P31' and these are separated by comma.

1. Special Handling Code (SCCCOD): Special Handling Code (SCC) is a three letter Code assigned by Air Cargo Tariff and Rules (TACT) for different cargo categories. Examples are PER (perishable foodstuff), AVI (live animal shipments), HUM (Human Remains) etc.
2. Screening Code (SCRCOD): There will be a Screening Code for each AWB shipment.

Using these three types of features, the similarity approaches have been deployed. The systems are described in the next section.

2.2 Similarity Approach

Two steps are followed to identify anomaly data from the Airline dataset. In the first step, a feature vector has been generated for each data point based on each AWB shipment OD-pair and each commodity code. There are different types of commodity code (e.g. 'CATDOG', 'CHEMICALS', 'CUTFLOWERS', 'PHARMA', etc.). In the next step, similarity approaches are applied to identify anomaly data for each AWB shipment OD-pair and each commodity code.

Each category of the features, an individual analysis has been done with the chargeable weight.

2.2.1 Feature Generation

For each numerical feature, Pearson's correlation has been calculated with the initial chargeable weight for each AWB OD-pair (shipment origin and destination pair) for a particular commodity item. If the correlation value is greater than 0.5 for a particular numerical feature with the initial chargeable weight, the feature has been selected for similarity approaches, otherwise the feature has been dropped. For example, the correlation value of 'GRSVOL' with respect to 'CHGWGT' is greater than 0.5, the 'GRSVOL' is considered as the feature for

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

similarity approaches. Before applying the Pearson's correlation, each of the numeric feature is normalised between [0,1].

For single valued categorical feature, an one-hot encoding transformation has been done and each of the individual category's variance within single valued categorical feature is being calculated with initial chargeable weight. If an individual category's variance is differ from overall variance with a certain threshold value (0.10), the individual category is considered as a feature to the similarity approaches. For example, in 'BOM_DEL' OD-pair and 'GENERAL' commodity code, the agent code of '1234567' is considered as an individual categorical feature because the chargeable weight variance of the agent code '1234567' is differ more than 0.10 with respect to overall variance.

For each of multi-valued categorical feature, an equal length vector has been constructed with the values of zero's and one's separated by comma. 'one' represents the presence of the individual category. 'zero' means the individual category is absent in multi-valued categorical feature. Each individual category's variance within multi-valued categorical feature is being calculated and compared with overall variance. If the individual category's variance is differ from a certain threshold, the individual feature is considered as a feature for the similarity approaches, otherwise the individual category will be dropped. For example, in 'BOM_DEL' OD-pair and 'CATDOG' commodity code, there are 11 unique 'SCCCOD' (AVI, BIG, P03, P06, P10, P13, P19, P31, P45, P54, P60). For a particular data point which is having 'SCCCOD' of AVI, BIG and P31, are coded as 1,1,0,0,0,0,0,1,0,0,0 where '1' represents the presence of individual component and '0' represents the absence of individual component. Later coded values are separated based on comma and individual separated component variance is being calculated and compared with overall variance. If the variance difference value is greater than 0.10, the individual component is selected as feature.

These three types of features are concatenated and made a feature vector for each data point based on each AWB shipment OD-pair and each commodity code. The feature vectors are passed to the similarity approaches to identify anomaly/non-anomaly data points. In the next section the descriptions of similarity approaches are mentioned.

2.2.2 Simple Similarity Approach (SSA)

Cosine similarity [7] is a metric used to determine how the data points are similar with respect to other data points using the feature vectors mentioned in the above section. Mathematically, the similarity approach measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this method, each test vector cosine score is computed with all the training vectors and reference training vector is identified based on the highest cosine score. The label of reference training vector is assigned to the test vector label.

For example, there are 21 training data points for the shipment OD-pair 'BOM_DEL' and commodity code 'CATDOG'. The labels of the training data points are 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 and 1 respectively. The feature vector (also called training vector) of each training data point is generated by using the concatenation of three different features mentioned in the above section. For a particular test data point for the same shipment OD-pair and same commodity code, a feature vector (also called test vector) is generated. Cosine scores of the test vector are calculated with all the 21 training vectors. The highest cosine score of the test data is getting with 17th training data point. The test data point label ('0') is assigned with the 17th training data label. The simple similarity model architecture is shown in Figure 1.

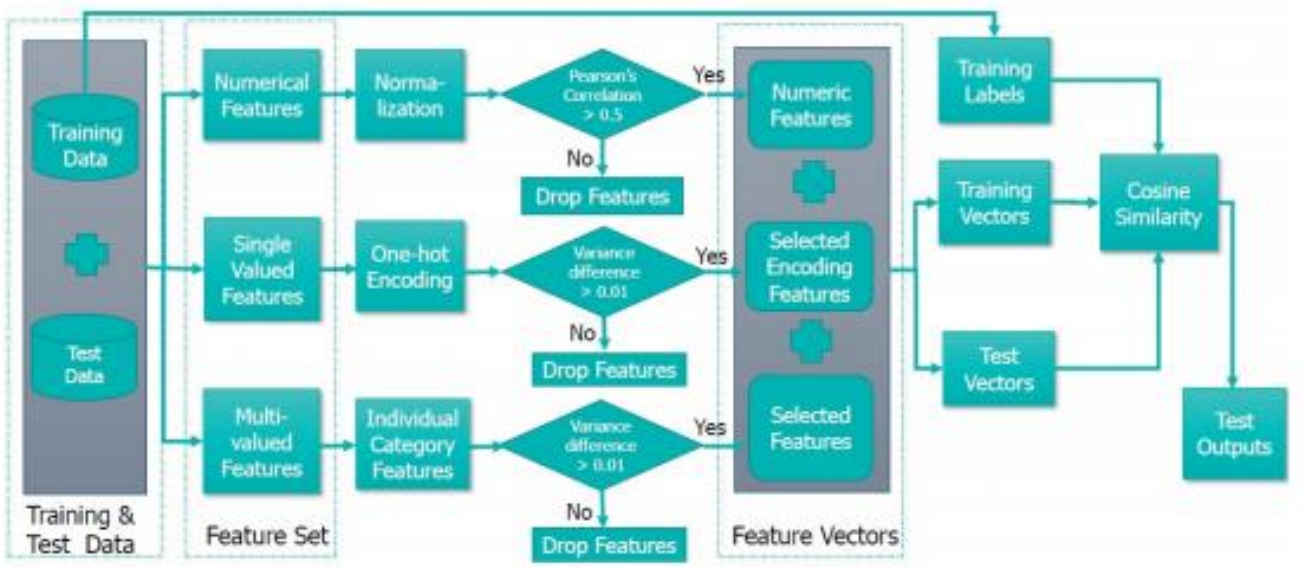

Figure 1: Simple Similarity Approach

2.2.3 Majority Voting Similarity Approach (MVSA)

A majority voting approach is applied on cosine similarity approach. In this approach, first top k training reference vectors are identified for each test vector based on highest cosine scores. Later the test vector label is assigned based on the majority voting of the top k training reference vectors' labels. Here the value of k is set to 5. For example, for the shipment OD-pair 'BOM_DEL' and commodity code 'CATDOG', the highest cosine values of the test vector are 0.9532, 0.9376, 0.8465, 0.8193 and 0.8011 with respect to the reference training vectors of 17th, 12th, 20th, 3rd and 6th, respectively mentioned the same example in the above Section 2.2.2. The labels of the reference training vectors are '0', '0', '0', '0' and '1', respectively. The number of times of '0' and '1' of the reference training vectors are 4 and 1 respectively. Based on the majority (4 times as '0' and 1 time as '1'), the test vector label is assigned to '0'.

2.2.4 Weighted Voting Similarity Approach (WVSA)

In this approach, first top k reference training vectors are identified for a test vector with respect to the highest cosine scores. A weight has been assigned for each of the reference training vector. More weight is given to the reference training vector if the cosine score between the test vector and reference training vector is more. Later weight values are multiplied with the cosine scores and summed up the values for each category of the labels. The test vector label is assigned based on highest summed up category's label.

For example, the highest cosine values of the test vector are 0.9532, 0.9376, 0.8465, 0.8193 and 0.8011 with respect to the reference training vectors of 17th, 12th, 20th, 3rd and 6th,, respectively. The weight vectors are 1.00, 0.90, 0.80, 0.70 and 0.60 for the reference training vectors of 17th, 12th, 20th, 3rd and 6th,, respectively. The labels of the reference training vectors are '0', '0', '0', '0' and '1', respectively. The summed up scores are calculated for label '0' and '1' mentioned below.

$Score_0 = 0.9532 * 1.00 + 0.9376 * 0.90 + 0.8465 * 0.80 + 0.8193 * 0.70 = 3.0477$

$Score_1 = 0.8011 * 0.60 = 0.4807$

Here $Score_0 > Score_1$, so the label of the test vector is assigned to '0'.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

3 Dataset and Experiment

The experiments were done based on the Airline datasets. When any changes between initial and final chargeable weights for a particular data point is found, the data point is considered as anomaly. Fifty AWB OD-pairs are chosen where the frequency of anomaly for each AWB shipment OD-pair is greater than 0. Currently the above similarity approaches are conducted on one year seven months data where last four months data are considered as test data and rest of the other months data are considered as training data.

Table 1: *Dataset: statistic of anomaly and non-anomaly*

Data	Anomaly	Non-anomaly
Training Data	3,896	25,774
Test Data	1,497	9,302
Total	5,393	35,076

The statistics of the datasets are shown in Table 1. The training dataset contains 3,896 and 25,774 number of anomaly and non-anomaly data points, respectively mentioned in Table 1. The total number of anomaly and non-anomaly in the test data are 1,497 and 9,302, respectively.

3.1 Result

The similarity approaches are applied on Airlines data. The systems are evaluated based on precision, recall and F-measure values mentioned below:

Precision = Correctly Identified Anomaly / System Identified Anomaly
Recall = Correctly Identified Anomaly / Total number of Actual Anomaly
F-measure = 2 * Precision * Recall / (Precision + Recall)

The results are shown in Table 2. The simple similarity approach outperformed over all other approaches. The best system, namely SSA produces the precision, recall and F-measure of 41.12%, 54.91% and 47.02%, respectively. Once the Majority Voting Similarity Approach (MVSA) is applied, the precision value is increased 3.84% over the SSA but the recall value is dropped 9.35%. The similar behavior is also observed when Weighted Voting Similarity Approach (WVSA) is applied on the Airline data. The precision, recall and F-measure of WVSA are 45.04%, 45.76% and 45.39%, respectively.

Table 2: *Results: Test data performance in percentage*

System	Total Records	Actual	Predicted	Correctly Identified	Precision	Recall	F-measure
SSA	10799	1497	1999	822	41.12%	54.91%	47.02%
MVSA	10799	1497	1517	682	44.96%	45.56%	45.26%
WVSA	10799	1497	1521	685	45.04%	45.76%	45.39%

4 Error Analysis

An error analysis has been done on the airline dataset. A confusion matrix has been drawn for Simple Similarity Approach (SSA) mentioned in Table 3. In the confusion matrix Table 3, it is observed that many anomaly data points are not able to identify correctly by the system. The reason may be the very few number of training samples are available in the training dataset. It is also observed that many test data points are wrongly identified as anomaly which are actually non-anomaly. A manual checking has been conducted for a few AWB shipment OD-pairs where the highest cosine similarities are very low and the system wrongly identified as anomaly based on simple similarly approach.

Table 3: *Test Data Confusion Matrix*

Predicted / Actual	Cosine Similarity (SSA)	
	Anomaly	Non-anomaly
Anomaly	822	675
Non-anomaly	1177	9622

5 Conclusion and Future Work

Several similarity approaches are developed on airline data. The best performance was obtained using Simple Similarity Approach (SSA). In term of precision value, weighted voting similarity approach performed better that SSA. To enhance the system performance, it would make potentially be good to incorporate new features. In future, an experiment needs to conduct by increasing the training samples with more AWB shipment OD-pairs. In future, automated weights need to identify for each of the reference training samples using evolutionary approach, namely Genetic Algorithms. Ensemble approach could be introduced based on the combined outputs from different models for better performance. Finally, more and other types of models could be generated using other classification algorithms.

References

[1] Jun Kang Chow and Zhaoyu Su and Jimmy Wu and Pin Siang Tan and Xin Mao and Yu-Hsing Wang, "Anomaly detection of defects on concrete structures with the convolutional autoencoder," *Adv. Eng. Informatics*, vol. 45, no. x, pp. 101-105, 2020

[2] Raghavendra Chalapathy and Sanjay Chawla, "Deep Learning for Anomaly Detection: A Survey," *CoRR*, vol. abs/1901.03407, eprint. 1901.03407, 2019

[3] Dimitry Gorinevsky and Bryan Matthews and Rodney Martin, "Aircraft Anomaly Detection using Performance Models Trained on Fleet Data," C*onference on Intelligent Data Understanding, {CIDU}, Boulder, CO, USA*, pp. 17-23, 2012

[4] L. Li and M. Gariel and R. Hansman and R. Palacios, "Anomaly detection in onboard-recorded flight data using cluster analysis," *IEEE/AIAA 30th Digital Avionics Systems Conference*, pp.1-11, 2011

[5] Puranik, Tejas and Mavris, Dimitri, "Anomaly Detection in General-Aviation Operations Using Energy Metrics and Flight-Data Records," *Journal of Aerospace Information Systems*, vol. 15, pp. 1-14, 2017

[6] D. Huang and D. Mu and L. Yang and X. Cai, "CoDetect: Financial Fraud Detection With Anomaly Feature Detection," *IEEE Access*, vol. 6, pp. 19161-19174, 2018

[7] Zhua, S. and Wua, J. and Xiongb, H. and Xiaa, G., "Scaling up Top-K Cosine Similarity Search," *Data & Knowledge Engineering*, vol. 70, pp. 60-83, 2011

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922