Research on An Ensemble Anomaly Detection Algorithm

Yupeng Wang*, Shibing Zhu and Changqing Li
Space Engineering University, Beijing 101416, China
18600113718@126.com

Abstract. Aiming at the problem that the applicability of single anomaly detection algorithm is not strong in aerospace experiment, an ensemble anomaly detection algorithm is proposed. This algorithm combines multiple machine algorithms and can obtain better detection performance than any other algorithm. Through comparison, k-NN, PCA and HBOS are selected. These three algorithms have fast calculation speed and different algorithm mechanisms, which can effectively process various data sets. This paper first introduces the basic concept of anomaly detection, then introduces and explains the three algorithms, then integrates the three algorithms, and introduces the voting mechanism to vote on whether the sample points are normal. Finally, the performance of the algorithm is tested through simulation experiments. Compared with a single algorithm, the ensemble algorithm has better performance in precision and accuracy.

1. Introduction
In the space field, spacecraft may face various complex system failures, structural damage, abnormal energy consumption and other problems during operation. These problems are fed back through the data recorded by sensors. However, in the process of test evaluation before spacecraft launch, these potential problems may be discovered by mining and analyzing the "seemingly normal" data in the test. The spacecraft fault database is established by combining the sensing parameters and corresponding test data of the fault problem. In the process of test identification, the test data and the data in the database are compared and judged through the anomaly detection algorithm, so as to find out the hidden trouble of the problem in advance. Due to the various types and weak characteristics of spacecraft sensing data and test data, and the limited detection capability of individual anomaly detection algorithms, it is necessary to adopt an anomaly detection algorithm with high integration, simple calculation and high efficiency. Based on this premise, this paper proposes an anomaly detection algorithm based on integration method.

2. Related Work
Anomaly detection techniques are generally divided into three categories[1]: supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection. Supervising anomaly detection, knowing the data of normal type and anomaly type, and carrying out anomaly detection on the new data after training the known data; Semi-supervised anomaly detection, which only knows normal types of data, and carries out anomaly detection on new data after data training; Unsupervised anomaly detection does not know the data of normal type and anomaly type in advance, but only carries out anomaly detection by comparing the differences between the data in the data set. This paper adopts the technology of supervised anomaly detection.
The main anomaly detection technologies include: k-NN[2], LOF[3], HBOS[4] for anomaly detection based on proximity, PCA[5], OCSVM[6] for anomaly detection based on linear model, ABOD[7] for anomaly detection based on probability, iforest[8] for anomaly detection based on integration method, and AutoEncoder[9] for anomaly detection based on neural network. This paper mainly integrates KNN, PCA and HBOS, so it introduces them.

2.1. k-NN
K-Nearest Neighbor Algorithm is a relatively classical algorithm. The distance between a point and its K nearest neighbors is calculated through the distance-based discrete point formula. Each point can be sorted according to the distance between each point and its nearest neighbors, and the vertices in the sorting can be regarded as outliers[2]. The k-NN algorithm is simple and easy to implement. It is suitable for many kinds of abnormal problems, but it is computationally expensive and has poor interpretability.

\[ D^k(A) \] is usually used to represent the distance from point a to its k nearest neighbor. Scoring is based on the distance of \( D^k(A) \). The definition of abnormal point \( D_n^k \) is as follows: Given a set of data sets, this set of data sets includes N points (including n abnormal points), and each point sets K neighboring points. If there are not more than n-1 remaining points a satisfying \( D^k(a) > D^k(A) \), then this is a \( D_n^k \) anomaly.

In short, if the points are arranged according to the distance of \( D^k(A) \), then the n points at the top are considered abnormal points. The distance of each point can be measured by Manhattan distance or Euclidean distance.

2.2. PCA
Anomaly detection through principal component analysis (PCA) is mainly realized by measuring the distance difference between abnormal instances and normal instances in principal component space. Anomaly detection by PCA can process high-dimensional data without hypothesis model, and the processing speed is very fast.

Data sets generally contain one or several unusual observations, which are different from the results of other observations after dimensional transformation. When the observation result is different from most data or unlikely to appear under the assumed data probability model, it is considered as an abnormal value. In high dimensions, there may be outliers in a single dimension. However, these values will not be analyzed separately, but all features will be considered by multiple methods.

Let \( X_1, X_2, \ldots, X_n \) be a random sample having an average vector \( \mu \) and a covariance matrix \( \Sigma \), where:

\[
X'_j = (X_{j1}, X_{j2}, \ldots, X_{jp}), \ j = 1, 2, \ldots, n
\]  

(1)

If the distribution of \( X_1, X_2, \ldots, X_n \) is multivariate normal distribution, T2 statistics based on Manhattan distance for an observation object x with the same distribution is:

\[
T^2 = \frac{n}{n+1}(X - \bar{X})'S^{-1}(X - \bar{X})
\]  

(2)

It is the distribution of \( \frac{(n-1)}{n-p}F_{p,n-p} \), where:

\[
\bar{X} = \frac{1}{n}\sum_{j=1}^{n}X'_j, \ S = \frac{1}{n-1}\sum_{j=1}^{n}(X'_j - \bar{X})(X'_j - \bar{X})'
\]  

(3)

\( F_{p,n-p} \) represents a random variable with p and n-p degrees of freedom in f distribution. When the value of \( T^2 \) is large, it indicates that the observed value \( X \) has a large deviation from the center of the whole population, and the statistical value \( \frac{(n-1)}{n-p}T^2 \) of the measured F distribution can be used to detect an outlier.

PCA is commonly used in multivariate anomaly testing, assuming \( y_1, y_2, \ldots, y_p \) are the sample principal components.

\[
y_i = e_i'(x - \bar{x}), i = 1, 2, \ldots, p
\]  

(4)

The normalized sum of squares of principal component fractions is equal to the Mahalanobis distance from the observed value \( x \) to the average value of the sample.
\[ \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_p^2}{\lambda_p} \]  (5)

Since the principal components of the samples are uncorrelated, it is necessary to have a chi-square distribution with degree of freedom \( q \) under the normal assumption and the assumption that the samples are large, and it is necessary to assume that all eigenvalues are positive and different.

\[ \sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q}, \quad q \leq p \]  (6)

The decision rule for outliers is:

Given the significance level value \( \alpha \), if \( \sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi^2_q(\alpha) \), then the sample is abnormal. Where \( \chi^2_q(\alpha) \) is the upper \( \alpha \) percentage point of chi-square distribution with degree of freedom \( q \), \( \alpha \) represents the error or false alarm probability when normal values are detected as abnormal values [5].

2.3. HBOS

HBOS (Histogram-based Outlier Score) is an anomaly detection algorithm based on histogram estimation density, which scores features in linear time to find out anomaly samples. It is based on the assumption that features are independent of each other, and its running speed is fast, but its accuracy is insufficient.

HBOS algorithm[4] step method is as follows:

Step1 First constructs histograms of individual variables. If it is data representing the category, simply count each category and calculate the relative frequency (i.e. the height of the histogram). When it is data representing numerical value, dynamic column width histogram is adopted to construct;

Step2 sorts the values, grouping a fixed number \( (N/k) \) of continuous values into a single rectangular column, where \( n \) is the total number of instances and \( k \) is the number of rectangular columns \( (k = \sqrt{N}) \);

Step3 The area of each rectangular column represents the observed quantity. From the first step, it can be seen that the areas of all rectangular columns are equal. Since the width of the rectangular columns is determined by the first and last values, the height of each rectangular column can be determined under the same area. Rectangular columns with larger coverage have smaller height, which indicates lower density.

Step4 normalizes the determined histogram to a maximum height of 1.0 to ensure that the weights of different features and outliers are equal;

Step5 calculates the HBOS of each sample \( p \) using the corresponding height of the matrix where the samples are located, which is equal to the reciprocal of the mutually independent feature estimation density:

\[ \text{HBOS}(p) = \sum_{i=0}^{d} \log \left( \frac{1}{\text{hist}_i(p)} \right) \]  (7)

3. Anomaly Detection Algorithm Based on Ensemble

Based on the idea of integration, it is to fuse various algorithms. In actual operation, it is to fuse all kinds of algorithms together to jointly solve a certain problem. These algorithms can be different algorithms or the same algorithm[10-11]. Its composing thought is: selecting a group of Individual Learner, and using some strategy to combine them, finally, solving specific problems.

This paper integrates KNN, PCA and HBOS algorithms, and determines whether a point is abnormal by voting. The characteristic of these three methods is that they run very fast and can find abnormal data in the data set almost in real time. Because different methods may have limitations, it is not possible to accurately judge all target objects, while integrating multiple algorithms can better and more comprehensively detect abnormal points. PCA is a linear model algorithm, HBOS and k-NN are probability-based algorithms. The combination of the three methods can achieve certain complementarity and can detect in a wider range of data types.

The specific steps of the algorithm are as follows:
Step 1 reads the data and divides the data into test data and training data, including characteristic parameter \( X (X') \) and result parameter \( y (y') \).

Step 2 combines the idea of integration and recursively uses three anomaly detection algorithms. Firstly, the training data set \( X' \) is used for training. After training, the test data \( X \) is calculated to obtain the anomaly score of the test data. Then, whether a sample is normal or not is determined according to the anomaly score to obtain the detection label.

Step 3 introduces a voting mechanism, and according to the discrimination results of the three algorithms, adopts the principle that the minority is subordinate to the majority, and discriminates the performance of a sample to determine whether it is normal at last, and obtains the abnormal \( label\_vote \) after voting:

Step 4 the confusion matrix of test data is obtained by calculating \( y \) and \( label\_vote \).

The pseudocode is as follows:

Algorithm VOTE_AD (Date)

Input: Date

Output: Confusion matrix

Step:
1: \( X, y, X', y' \) ← data
2: for AnomalyDetection do
3:   fit \( (X') \)
4:   \( score \) ← DecisionFunction \( (X) \)
5:   \( label \) ← \( score \)
6: end for
7: \( label\_vote \) ← vote(\( label \))
8: Return Confusion matrix(\( y, label\_vote \))

4. Experiments and Results

In order to verify the algorithm proposed in this paper, five real data sets of Letter, Mnist, Musk, Pendigits and Satellite provided by PYOD toolbox are used to simulate the anomaly detection algorithm under 64-bit windows10 operating system and python3.7 compilation environment. The evaluation of anomaly detection in this paper is to obtain 4 kinds of judgment forms of each data set in the confusion matrix, and further calculate the accuracy, precision and recall of different data sets[12]. Table 1 shows all kinds of experimental data, and Tables 2 to 6 show experimental results.

Table 1. All kinds of data tested.

| Data   | Sample | Dimensions | Outlier Perc |
|--------|--------|------------|--------------|
| Letter | 1600   | 32         | 6.25         |
| Mnist  | 7603   | 100        | 9.21         |
| Musk   | 3062   | 166        | 3.17         |
| Pendigits | 6870 | 16         | 2.27         |
| Satellite | 6435 | 36         | 31.64        |

Table 2. Letter dataset experimental results.

| Letter | KNN | PCA | HBOS | VOTE_AD |
|--------|-----|-----|------|---------|
| TP     | 552 | 541 | 540  | 555     |
| FN     | 47  | 58  | 59   | 44      |
| FP     | 24  | 35  | 36   | 35      |
| TN     | 17  | 6   | 5    | 6       |
| Accuracy | 0.8891 | 0.8547 | 0.8516 | 0.8766 |
| Precision | 0.9583 | 0.9392 | 0.9375 | 0.9407 |
| Recall  | 0.9215 | 0.9032 | 0.9015 | 0.9265 |
Table 3. Mnist dataset experimental results.

|       | KNN | PCA | HBOS | VOTE_AD |
|-------|-----|-----|------|---------|
| TP    | 2586| 2588| 2505 | 2603    |
| FN    | 177 | 175 | 258  | 160     |
| FP    | 151 | 149 | 232  | 162     |
| TN    | 128 | 130 | 47   | 117     |
| Accuracy | 0.8922 | 0.8935 | 0.8389 | 0.8941 |
| Precision | 0.9448 | 0.9456 | 0.9152 | 0.9414 |
| Recall | 0.9359 | 0.9367 | 0.9066 | 0.9421 |

Table 4. Musk dataset experimental results.

|       | KNN | PCA | HBOS | VOTE_AD |
|-------|-----|-----|------|---------|
| TP    | 1078| 1102| 1102 | 1131    |
| FN    | 111 | 87  | 87   | 58      |
| FP    | 24  | 0   | 0    | 0       |
| TN    | 12  | 36  | 36   | 36      |
| Accuracy | 0.8898 | 0.9290 | 0.9290 | 0.9527 |
| Precision | 0.9782 | 1.0000 | 1.0000 | 1.0000 |
| Recall | 0.9066 | 0.9268 | 0.9268 | 0.9512 |

Table 5. Pendigits dataset experimental results.

|       | KNN | PCA | HBOS | VOTE_AD |
|-------|-----|-----|------|---------|
| TP    | 2435| 2453| 2459 | 2527    |
| FN    | 258 | 240 | 234  | 166     |
| FP    | 38  | 20  | 14   | 22      |
| TN    | 17  | 35  | 41   | 33      |
| Accuracy | 0.8923 | 0.9054 | 0.9098 | 0.9316 |
| Precision | 0.9846 | 0.9919 | 0.9943 | 0.9914 |
| Recall | 0.9042 | 0.9109 | 0.9131 | 0.9384 |

Table 6. Satellite dataset experimental results.

|       | KNN | PCA | HBOS | VOTE_AD |
|-------|-----|-----|------|---------|
| TP    | 1687| 1745| 1743 | 1745    |
| FN    | 70  | 12  | 14   | 12      |
| FP    | 629 | 571 | 573  | 570     |
| TN    | 188 | 246 | 244  | 247     |
| Accuracy | 0.7284 | 0.7735 | 0.7720 | 0.7739 |
| Precision | 0.7284 | 0.7535 | 0.7526 | 0.7538 |
| Recall | 0.9602 | 0.9932 | 0.9920 | 0.9932 |

By comparing the experimental results of the five data sets, the following conclusions can be drawn:

1. From the point of view of accuracy, recall and precision, the integration algorithm does have certain improvement compared with other separate algorithms, especially the comparison of precision, the integration algorithm ranks first. In terms of accuracy, except for the second place in letter data set, the rest are ranked first. The improvement of recall rate is relatively weak, with two data sets in the first place and three data sets in the second and third places in the middle. However, the integration algorithm and the algorithm in the first place are relatively similar.

2. Accurate to the number of normal samples and abnormal samples, the integration algorithm is very effective in the detection of normal samples, finding the largest number of normal samples. However, the ability to detect abnormal samples is weak. There are two data sets with the largest
number of abnormal samples found and three in the middle, which corresponds to the recall rate. To a certain extent, this shows that normal data have certain rules and small individual differences. By integrating various algorithms, more points can be detected, while the differences between abnormal data are large. Different algorithms also have large differences in the detection of characteristic attributes of abnormal samples, so it is not easy to distinguish.

3. One data set that needs to be noticed is Musk. Compared with the other four data sets, musk has more obvious abnormal features, so various algorithms have better detection effect on abnormal features in the data. In addition, under the condition that the algorithm can correctly judge all normal samples (i.e. recall ratio is 1), the performance index results of accuracy and precision ratio of the integrated algorithm are greatly improved.

4. The integration of anomaly detection is different from the integration of classification, because anomaly detection pays more attention to anomaly samples. In the classification process, it is mainly to distinguish as many normal samples as possible, while ignoring samples other than normal samples. In addition, different types of anomalies may have different characteristics. If classified, the result may be a large set containing normal data and many small sets containing abnormal data. In general, when classified, several large sets belonging to different categories are often obtained. Compared with abnormal data, the characteristics of the data of the same category in the classification are more consistent. Therefore, the classification based on integration obtains better results, while the anomaly detection based on integration needs to be improved in algorithm.

5. Summary and Future Work
In view of the limitation of single anomaly detection algorithm, this paper proposes an anomaly detection algorithm based on integration, which can solve the defect of performance difference of single algorithm on different data sets. Through experiments on the selected five data sets, the algorithm improves the calculation of relevant performance indexes to a certain extent, but the detection ability of actual abnormal samples needs to be improved. The next research focus is to continue to optimize the algorithm, make it more sensitive to abnormal samples, and be able to detect more abnormal points, thus improving the performance of anomaly detection.

References
[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3):1-58.
[2] Angiulli F, Pizzuti C. Fast Outlier Detection in High Dimensional Spaces[C]//Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002. Springer-Verlag, 2002.
[3] Sander J. LOF: Identifying Density-Based Local Outliers[J]. ACM SIGMOD Record, 2000, 29(2):93-104.
[4] Goldstein M, Dengel A. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm[C]//KI-2012: Poster & Demo Track, 2012.
[5] Shyu M L, Chen S C, Sarinnapakorn K, et al. Principal component-based anomaly detection scheme[M]. Berlin: Springer, 2006.
[6] Scholkopf B, Platt J C, Shawetaylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2014, 13(7):1443-1471.
[7] Kriegel H P, Hubert M S, Zimek A. Angle-based outlier detection in high-dimensional data[C]//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2008.
[8] Liu F T, Kai M T, Zhou Z H. Isolation Forest[C]//Eighth IEEE International Conference on Data Mining. 2008.
[9] Chen J, Sathe S, Aggarwal C, et al. Outlier detection with autoencoder ensembles[C]//SIAM International Conference on Data Mining, 2017.
[10] Z H Zhou. Ensemble Methods Foundations and Algorithms[M]. CRC Press, 2012.
[11] Campos G O, Zimek A, Meira W. An Unsupervised Boosting Strategy for Outlier Detection Ensembles[C]//Pacific-Asia Conference on Knowledge Discovery & Data Mining, 2018.

[12] AGGARWAL C C. Outlier Analysis[M]. 2nd ed. Berlin: Springer, 2016.