IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 33, No. 4, August, 2019, pp. 321-325
Journal homepage: http://iieta.org/journals/ria

Detection of Abnormal Oil Data Based on Feature Selection

Bixin Li[1], Rong Fan[1*], Bing Yang[2], Shigang Lin[1]

[1] Department of Petroleum, Army Logistics University of PLA, Chongqing 401311, China
[2] Army Logistics Department of PLA, Beijing 100072, China

Corresponding Author Email: frfr1002@126.com

https://doi.org/10.18280/ria.330409

Received: 1 March 2019
Accepted: 14 July 2019

Keywords:
abnormal oil data, oil management, feature selection, fisher score

ABSTRACT

To achieve effective oil management, it is critical to disclose the laws of oil supply, consumption, and natural loss through data analysis. However, the accuracy of data analysis is often suppressed by the mistakes and irrelevance of the input data, which are inevitable due to the large size and diversity of the data collected from the oil depots. To solve the problem, this paper proposes an abnormal oil data detection approach based on feature selection (AODDFS). In the AODDFS, the format of the input data was preprocessed to satisfy the requirements of feature selection; the fisher score was then employed to compute the relevance of each entry with normal features; finally, the abnormal entries were located based on the relevance values. Then, the AODDFS results were analyzed with boxplot and standard deviation. Finally, the AODDFS was verified through a case study on the data collected from several large oil depots. The results show that the AODDFS can effectively detect abnormal oil data with a precision of 85.00% and a recall of 80.94%.

1. INTRODUCTION

The key to scientific and reasonable oil management [1-4] include identifying the most typical oil depots, collecting the data from these depots, and disclosing the laws of oil supply, consumption, and natural loss through data analysis [5-10]. With the rapid development of information technology, many computing methods suitable for oil management have emerged. The effectiveness of these methods depends on the quality of the input data, i.e. the data on oil supply, consumption, and natural loss [11, 12]. However, it is very difficult to collect a reliable set of input data, because they exist in large quantities and with high diversity [13].

Many errors are likely to occur during the collection of the massive and diverse data, such as wrong entries, irrelevant entries and typos. If inputted to the computing methods, the erroneous data will affect the accuracy of the data analysis and the ensuing decisions on oil management, and are thus called the abnormal oil data [14]. This type of data must be eliminated to ensure the reliability of the input data. Nevertheless, it takes a long time to detect the abnormal oil data out of the huge and diverse input data. What is worse, the existing detection methods mainly rely on manual checking (e.g. setting abnormal intervals and making empirical judgements). This calls for an automated detection approach that can deal with large and various datasets.

Recently, feature selection, a.k.a. variable selection or attribute selection, has attracted much attention for its ability to filter out abnormal oil data [15, 16]. Feature selection mainly identifies a subset of relevant features (e.g. variables and predictors) from the original data, facilitating data analysis, model construction and abnormal feature detection. This technique has often been adopted to pinpoint and remove useless, irrelevant or redundant attributes from the original data. An attribute is considered useless, irrelevant or redundant if it does not contribute and even suppress the accuracy of the prediction model [17].

Feature selection has been proved as an effective tool to remove abnormal features without sacrificing the prediction performance [18, 19]. The existing feature selection methods generally fall into three categories: the filter methods, the wrapper methods, and the embedded methods. Specifically, the filter methods evaluate each feature against the general characteristics of data, without using any learning algorithm; the wrapper methods select features according to the performance of a preset learning algorithm; the embedded methods identify the target features in the training process, and analyze feature relevance based on the objective of the customized learning model. The wrapper methods usually produce a subset of selected features. Meanwhile, the filter methods and embedded methods either output the relevance scores of all features or generate a subset of selected features. Depending on the type of output, the latter two kinds of methods can also be categorized into feature scoring algorithms and subset selection algorithms.

As stated above, feature selection can identify the irrelevant attributes in the original data. Irrelevant attributes are equivalent to abnormal data, because their features differ from those of relevant attributes. Taking each data entry as a feature, the abnormal oil data can be detected naturally from the set of input data through feature selection. Following this train of thought, this paper proposes an abnormal oil data detection approach based on feature selection (AODDFS). The most popular filter method, fisher score [20, 21], was selected as the feature selection tool to detect abnormal oil data. The filter method was adopted because it does not require any learning algorithm like wrapper and embedded methods. In the AODDFS, the format of the input data was preprocessed to satisfy the requirements of feature selection; the fisher score was then employed to compute the relevance of each entry

with normal features; finally, the abnormal entries were located based on the relevance values. Then, the AODDFS results were analyzed with boxplot and standard deviation. Finally, the AODDFS was verified through a case study on the data collected from several large oil depots. The results show that the AODDFS can effectively detect abnormal oil data with a precision of 85.00% and a recall of 80.94%.

The contribution of this paper can be summarized as follows:

(1) We propose an abnormal oil data detection approach by using feature selection.

(2) We conduct a case study to evaluate our approach, showing that our approach is effective to detect abnormal oil data.

(3) We demonstrate the potential of a perspective of using feature selection for effective abnormal oil data detection.

The remainder of this paper is organized as below. Section 2 describes abnormal oil data detection based on fisher score. Section 3 presents the analytical tools of AODDFS results. Section 4 shows the case study and Section 5 draws the conclusion.

2. ABNORMAL OIL DATA DETECTION BASED ON FISHER SCORE

This section details our abnormal oil data detection approach AODDFS using fisher score. In recent years, fisher score has been increasingly used as an effective feature extractor for irrelevant data detection and classification [20, 21]. The main idea of fisher score is to remove the irrelevant features from the original data, creating a subset of selected features. In the data space extended from the selected features, the distances between the data points in the same class are minimized, while those between the data points in different classes are maximized [21]. By this method, the relevance of each feature in each entry is evaluated independently against the score obtained under the fisher criteria. The higher the score, the greater the relevance of the feature, and the inverse is also true. The features with low relevance belong to abnormal oil data in the field of oil management. Below is the mathematical description of the fisher score-based abnormal oil data detection.

For a given dataset $\{(x_i, y_i)\}_{i=1}^{n}$ ($x_i \in R^m$; $y_i \in \{1,2,\ldots,c\}$), the set of input data can be expressed as a matrix $X = [x_1, x_2, \ldots, x_n] \in R^{m \times n}$, where x^j is the oil data entry in the j-th row, and y_i is the class label of x_i. Let 1 be a vector of all ones with a proper length, 0 be a vector of all zeros, and I be an identity matrix with a proper size.

Taking each entry as a feature, the input matrix can be considered to have m features. In this way, the detection of abnormal oil data can be transformed into a feature selection problem, that is, abnormal oil data are equivalent to highly irrelevant features. If d features are selected from the m features, the input matrix $X \in R^{m \times n}$ will be reduced to $Z \in R^{d \times n}$. Then, the fisher score can be defined as:

$$F(Z) = tr\{(S_b^{'})(S_t^{'} + \gamma I)^{-1}\} \tag{1}$$

where, γ is a positive regularization parameter; $S_b^{'}$ is the between-class scatter matrix; $S_t^{'}$ is the total scatter matrix; γI is a perturbation term to make the normally singular $S_t^{'}$ being positive semi-definite. The two scatter matrices can be computed by:

$$\begin{aligned} S_b^{'} &= \sum_{k=1}^{c} n_k (\mu_k^{'} - \mu^{'})(\mu_k^{'} - \mu^{'})^T \\ S_t^{'} &= \sum_{i=1}^{n} (z_i - \mu^{'})(z_i - \mu^{'})^T \end{aligned} \tag{2}$$

where, $\mu_k^{'}$ and n_k are the mean vector and size of the k-th class in the reduced data space Z, respectively; $\mu^{'} = \sum_{k=1}^{c} n_k \mu_k^{'}$ is the overall mean vector of the reduced data.

Since there are $\binom{m}{d}$ candidate Z s out of X, the feature selection is a very difficult problem of combinatorial optimization. To overcome the difficulty, the commonly used heuristic strategy was adopted because our feature selection model does not concern the number of reduced features. Specifically, a score was computed independently for each feature under criterion F, that is, there are only $\binom{m}{1} = m$ candidates. Thus, μ_k^j and σ_k^j are defined as the mean and standard deviations of the k-th class that the j-th feature belongs to, and μ^j and σ^j are defined as the mean and standard deviations of the whole dataset in accordance to the j-th feature. Then, the fisher score of the j-th feature can be defined as:

$$F(x^j) = \frac{\sum_{k=1}^{c} n_k (\mu_k^j - \mu^j)^2}{(\sigma^j)^2} \tag{3}$$

where, $(\sigma^j)^2 = \sum_{k=1}^{c} n_k (\sigma_k^j)^2$. The fisher score of each feature can be computed by equation (3). After the computation, the features with very low scores can be identified as abnormal ones, and each abnormal feature represents an abnormal oil data entry. Figure 1 gives an example of the fisher score-based detection of abnormal oil data.

	x1	x2	x3	x4	x5	x6	
	Date (Y/M/D)	Oil Depth (mm)	Oil Temprature (℃)	Apprarent Density (kg/m^3)	Standard Volume (m3)	Weight (kg)	
Entry_1	20090109	2,172	0.6	807.2	51,644	42,366	x^1
Entry_2	20090109	2,183	2.8	807.2	52,159	42,703	x^2
Entry_3	20090109	2,315	3.2	807.2	60,141	49,216	x^3
Entry_4	20090109	2,192	3.2	807.2	52,509	42,970	x^4
Entry_5	20090109	2,183	3.3	807.2	**231,234**	42,870	x^5
	y_1	y_2	y_3	y_4	y_5	y_6	
	1	2	2	2	2	2	

Figure 1. An example of the AODDFS

As shown in Figure 1, the task is to judge if the five oil data entries (features) are normal. In the example, the standard volume of Entry_5 was incorrect (231,234 in red) and should be 52,386. Since "Date" differs greatly from the other attributes about oil parameters, the class label *1* was given to "Date" and *2* was given to the other attributes. Using fisher score method, the scores of the five features were computed as (Entry_1, 129815.00), (Entry_2, 127424.91), (Entry_4, 125754.92), (Entry_3, 112671.39) and (Entry_5, 8371.51). Entry_5 is obviously an abnormal oil data entry, because it has a much smaller score (8,371.51) than the other features. This data entry must be corrected or removed from the input data. Otherwise, it will exert a negative impact on the data analysis.

3. ANALYTICAL TOOLS OF AODDFS RESULTS

By the AODDFS, each entry is assigned a score about its normality, and an entry with a low score is likely to be abnormal. However, the oil managers cannot identify an abnormal entry rapidly, if its score is low but not very low. Hence, the boxplot and standard deviation were employed to analyze the AODDFS results in visual and non-visual manners, respectively.

3.1 Boxplot

In descriptive statistics, a boxplot [22] clearly depicts a group of numerical data with their quartiles. It is a standard way to display the data distribution based on the Five-Number Summary: minimum, first quartile, median, third quartile, and maximum. In the simplest box plot (Figure 2), the central rectangle spans from the first quartile to the third quartile. The span is called the interquartile range (IQR). The median is indicated by the segment in the rectangle, and the minimum and maximum are above and below the box, respectively.

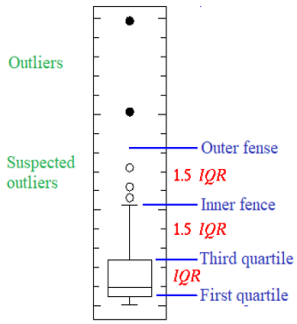

Figure 2. Boxplot structure

The interval between the minimum and the maximum is the full range of variation, the IQR is the likely range of variation, and the median is a typical value. The real datasets with surprisingly high maximums or surprisingly low minimums called outliers or suspected outliers. The outliers are either 3×IQR or more above the third quartile or 3×IQR or more below the first quartile. The suspected outliers are slightly more central versions of outliers: either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile.

In this paper, the boxplot is utilized to display the AODDFS results, making it easier for oil managers to look for abnormal entries. Since a high score means high normality, the managers only need to focus on the entries in suspected outliers and outliers of low scores. Besides, the boxplot provides the managers with a panorama of the distribution of all scores, enabling them to set a lower limit for entry check. Note that the suspected outliers were taken as the standard to check the abnormal scores, that is, the low scores belonging to suspected outliers.

3.2 Standard deviation

In statistics, the standard deviation, denoted as σ, is a measure of the variation or dispersion of a dataset. A low standard deviation indicates that the data points are close to the mean of the dataset, while a high standard deviation means the data points are scattered across a wide range. The standard deviation is defined as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4}$$

where, $X=\{x_1, x_2, ..., x_N\}$ is a dataset with N entries; $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the mean value of the dataset.

Based on standard deviation σ, a confidence interval $[\mu+\sigma,\mu-\sigma]$ was defined for entry check. More attention should be paid to the scores outside this interval. Since a high score means high normality, the oil managers only need to emphasize on the entries whose scores are below the interval, i.e. lower than $(\mu-\sigma)$.

4. CASE STUDY

4.1 Data collection

To verify our approach AODDFS in realistic scenarios, the input data were collected from several large oil depots in Beijing, China. Each of these oil depots supplies oil to smaller depots in Beijing and nearby provinces. The collected data cover the parameters of oil transmitted from one depot to another. Two datasets were selected for analysis, namely, Dataset_1 and Dataset_2. The former has 20,000 entries and the latter has 1,992 entries. There are 11 abnormal entries (e.g. wrong number, typo, and non-transmission entry) in Dataset_1 and 10 in Dataset_2. The AODDFS results were processed and displayed by boxplot and standard deviation. The entire verification was performed on MATLAB 7.8.0.

4.2 Evaluation metrics

The AODDFS performance was evaluated by two metrics: precision and recall. Precision is the fraction of retrieved entries that are abnormal, and recall is the fraction of abnormal entries that are retrieved. The two metrics can be expressed as:

$$\begin{aligned} precision &= TA/(TA+FA) \\ recall &= TA/(TA+TN) \end{aligned} \tag{4}$$

where, TA is the number of true abnormal entries detected by the AODDFS; FA is the number of false abnormal entries detected by the AODDFS; TN is the number of true abnormal entries not detected by the AODFS. Thus, (TA+FA) is the total number of entries judged by the AODDFS as abnormal, and (TA+TN) is the total number of true abnormal entries in the dataset.

Therefore, precision reflects the usefulness of the detection results while recall demonstrates the completeness of the results. In simple terms, high precision means that the AODDFS returns many truer abnormal entries than false ones, while high recall means that the AODDFS returns most of the abnormal entries in the input dataset. The higher the precision or recall, the better the AODDFS performs.

4.3 Results analysis

The AODDFS was applied to detect the abnormal entries in each of the two datasets, yielding a score on the normality of each entry. Then, the scores were separately processed and displayed by boxplot and standard deviation.

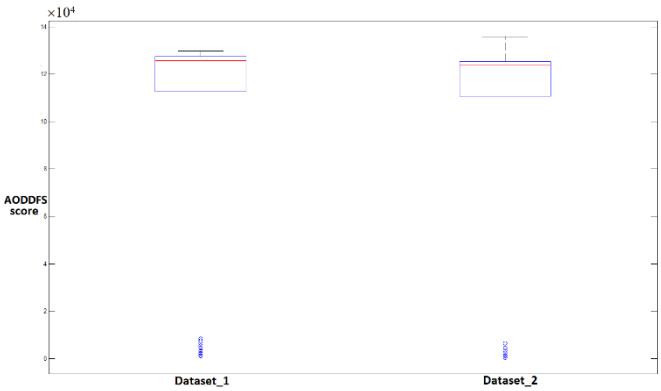

Figure 3. Boxplots of two datasets

The boxplots of the two datasets (Figure 3) were constructed with the AODDFS scores as the inputs. As shown in Figure 3, several entries with low scores were suspected outliers. By the definition of suspected outliers, oil managers checked the entries whose scores are 1.5×IQR or more below the first quartile. For Dataset_1, 10 out of the 12 entries, which were judged by the AODDFS as abnormal, were true abnormal entries. For Dataset_2, 7 out of the 8 entries, which were judged by the AODDFS as abnormal, were true abnormal entries. Hence, the AODDFS achieved the precision and recall of 83.33% and 90.91% in Dataset_1, respectively, and 87.5% and 70% in Dataset_2, respectively.

The AODDFS scores were also analyzed by standard deviation. For Dataset_1, the mean score was 122,496.62, and the standard deviation was 7,789.11. The entries with scores below (122,496.62-7,789.11) =114,707.51 were abnormal. Hence, 10 out of the 12 entries, which were judged by the AODDFS as abnormal, were truly abnormal. For Dataset_2, the mean score was 120,603.98, and the standard deviation was 7,646.37. The entries with scores below (120,603.98-7,646.37) = 112,957.61 were abnormal. Therefore, 7 out of the 10 entries, which were judged by the AODDFS as abnormal, were truly abnormal. To sum up, the AODDFS achieved the precision and recall of 83.33% and 90.91% in Dataset_1, respectively, and 87.5% and 70% in Dataset_2, respectively. The analysis results by standard deviation agree well with those of the boxplot analysis, indicating that both are effective in evaluating the AODDFS results.

The abnormal oil data detected by the AODDFS are summed up in Figure 4, where TA, FA and FN are 17, 3, and 4 respectively. This means the AODDFS considered 20 entries as abnormal, among which 17 were truly abnormal. The 3 false abnormal ones were normal, although their scores were much larger or smaller than those of the other entries. This situation is not uncommon in practice. For example, the natural loss of oil is much higher than regular transmission, if a depot needs to supply oil to a remote place with no nearby depots. Besides, the two selected datasets have a total of 21 abnormal entries. Four of them, which are correlated, were not identified by the AODDFS. This is because each feature was analyzed independently by fisher score, with the aim to save computing cost. Based on the data in Figure 4, the final precision and recall of the AODDFS were computed as 85.00% and 80.94%, respectively. Thus, most of the entries detected by the AODDFS are truly abnormal, and most of the abnormal entries in the datasets were detected by the AODDFS. It is safety to say that the AODDFS is an effective way to detect abnormal oil data.

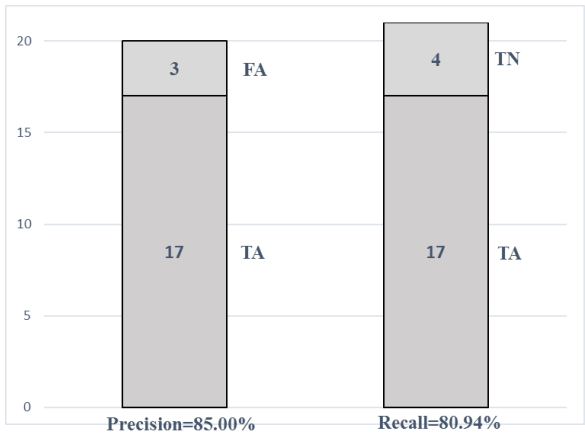

Figure 4. Summary of AODDFS results

5. CONCLUSIONS

In this paper, a practical method, the AODDFS, is proposed to detect the abnormal oil data based on feature selection. The fisher score was adopted to compute the relevance of each entry with normal features, and the AODDFS results were analyzed by boxplot and standard deviation. The experiments on real-world datasets show that the AODDFS achieved high precision and recall, revealing its effectiveness in detecting abnormal oil data. The future research will further optimize the AODDFS based on the relevance of dependent attributes, and include restriction rules to deal with rare cases.

REFERENCE

[1] El-Reedy, M.A. (2016). Project Management in the Oil and Gas Industry. Wiley, 297-313. https://doi.org/10.1002/9781119084129

[2] Sabri, H.A.R., Rahim, A.R.A., Yew, W.K., Ismail, S. (2017). An overview of turbomachinery project in Malaysian oil and gas industry. In IOP Conference Series: Materials Science and Engineering, 277(1): 1-9. https://doi.org/10.1088/1757-899X/277/1/012033.

[3] Rodhi, N.N., Anwar, N., Wiguna, I.P.A. (2017). A review on risk factors in the project of oil and gas industry. The Journal for Technology and Science, 28(3): 63-67. https://doi.org/10.12962/j20882033.v28i3.3217

[4] Pelitli, V., Doğan, Ö., Köroğlu, H.J. (2017). Waste oil management: Analyses of waste oils from vehicle crankcases and gearboxes. Global Journal of Environmental Science and Management, 3(1): 11-20. https://doi.org/10.22034/GJESM.2017.03.01.002

[5] Prates, A., Freigedo, F.E.A., Almeida, P.O. (2013). A critical assessment of the main challenges related to feasibility studies, risk analysis and monitoring of current offshore projects in Brazil. In Proceedings of Offshore Technology Conference, 29-31. https://doi.org/10.4043/24421-MS

[6] Safra, E.B., Antelo, S.B. (2010). Integrated project management applied in word-class gas-field development projects: From theory to practice. SPE Latin American and Caribbean Petroleum Engineering Conference Proceedings, 2: 1666-1676. https://doi.org/10.2118/139369-MS

[7] Sterling, G.H. (2013). Managing offshore megaprojects: Success is an option. Proceedings-SPE Annual Technical Conference and Exhibition, 4: 2773-2783. https://doi.org/10.2118/166310-MS
[8] Van Thuyet, N. (2007). Risk management in oil and gas construction projects in Vietnam. International Journal of Energy Sector Management, 1(2): 175-194. https://doi.org/10.1108/17506220710761582
[9] Enhassi, A. (2008). Risk management in building projects: owners' perspective. The Islamic University Journal, 16(1): 95-123. https://doi.org/10.5772/51460
[10] Anifowose, B., Lawler, D.M., der Horst, D.V., Chapman, L. (2011). Attacks on oil transport pipelines in Nigeria: A quantitative exploration and possible explanation of observed pattern. Applied Geography, 32(2): 636-651. https://doi.org/10.1016/j.apgeog.2011.07.012
[11] Namian, M., Albert, A., Zuluaga, C.M., Jaselskis, E.J. (2016). Improving hazard-recognition performance and safety training outcomes: Integrating strategies for training transfer. Journal of Construction Engineering and Management, 142(10): 04016048. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001160
[12] Osipova, E., Eriksson, P.E. (2013). Balancing control and flexibility in joint risk management: Lessons learned from two construction projects. International Journal of Project Management, 31(3): 391-399. https://doi.org/10.1016/j.ijproman.2012.09.007
[13] Corvellec, H. (2009). The practice of risk management: Silence is not absence. Risk Management, 11(3-4): 285-304. https://doi.org/10.1057/rm.2009.12
[14] Ghadge, A.T., Dani, S., Chester, M., Kalawsky, R. (2013). A systems approach for modelling supply chain risks. Supply Chain Management: An International Journal, 18(5): 523-538. https://doi.org/10.1108/SCM-11-2012-0366
[15] Duda, P., Stork, D.G. (2001). Pattern Classification. Wiley-Interscience Publication, 1-30. https://doi.org/10.1007/s00357-007-0015-9
[16] Guyon, I., Elisseeff, A., Kaelbling, L.P. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(7-8): 1157-1182. http://dx.doi.org/10.1162/153244303322753616
[17] Liu, H., Hiroshi, M. (2012). Feature selection for knowledge discovery and data mining. Springer Science & Business Media, 17-41. https://doi.org/10.1007/978-1-4615-5689-3
[18] Donoho, D.L. (2006). For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics, 59(7): 907-934. https://doi.org/10.1002/cpa.20131
[19] Ng, A.Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings, Twenty-First International Conference on Machine Learning, pp. 615-622. https://doi.org/10.1145/1015330.1015435
[20] Gu, Q., Li, Z., Han, J. (2011). Generalized fisher score for feature selection. Proceeding-UAI'11 Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pp. 266-273.
[21] Ahn, S., Korattikara, A., Welling, M. (2012). Bayesian posterior sampling via stochastic gradient fisher scoring. Proceedings of Twenty-Ninth International Conference Machine Learning, 2: 1591-159.
[22] Hubert, M., Vandervieren, E. (2008). An adjusted boxplot for skewed distribution. Computational Statistics and Data Analysis, 52(12): 5186-5201. https://doi.org/10.1016/j.csda.2007.11.008