Measurement 125 (2018) 48–55

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Natural gas pipeline valve leakage rate estimation via factor and cluster analysis of acoustic emissions

Shen-Bin Zhu[a], Zhen-Lin Li[a,*], Shi-Min Zhang[a], Le-Le Liang[a], Hai-Feng Zhang[b]

[a] *College of Mechanical and Transportation Engineering, China University of Petroleum-Beijing, 102249 Beijing, PR China*
[b] *PetroChina Pipeline R&D Center, 065000 Langfang, PR China*

ARTICLE INFO

Keywords:
Valve leakage recognition
Acoustic emission
K-means algorithm
K-medoids algorithm
Fuzzy C means algorithm
Factor analysis

ABSTRACT

This paper estimates the leakage rate of a valve in a natural gas pipeline via factor and cluster analysis of acoustic emission signals. Factor analysis was used to reduce the amount of redundant information in the highly dimensional features and extract the optimal features for the cluster analysis. Three types of clustering algorithm—fuzzy C means, *k*-means and *k*-medoids—were used to classify leakage rates. Performance was evaluated in terms of overall accuracy, computational time, iterations, Jaccard coefficients and Cohen's kappa. A model based on factor analysis and *k*-medoids clustering was found to be exceedingly effective for recognizing internal valve leakage rates. This method proved to be superior to the *k*-means and fuzzy C means clustering methods, and has potential value in real-world applications.

1. Introduction

During the operation of natural gas pipeline stations, internal valve leakage can lead to huge economic losses, environmental pollution and even accidental explosions that endanger the safety of personnel. As a means of unconventional non-destructive testing, acoustic emission (AE) testing has the characteristics of being highly efficient, fast and economical. Therefore, it has been widely used in the petroleum industry as a way to detect problems such as underground pipeline leakage [1,2], valve leakage [3] and atmospheric tank bottom leakage [4]. The feasibility and effectiveness of AE technology for detection of internal valve leakage has also received researchers' attention [3,5–7,16]. In addition, in comparison with the detection technologies of vibration analysis [8], infrared thermography [9,10] and dynamic pressure [11], AE technology is more stable and reliable in critical conditions such as those with high noise, low pressure and low leakage rates [6].

In the process of valve leakage detection, researchers are not only concerned about whether valve leakage occurs or not, but are more concerned with the valve leakage rate. Prediction of leakage rate provides the information necessary for valve maintenance. At present, the qualitative analysis of internal valve leakage in natural gas pipelines has reached maturity, but its quantitative analysis still requires further research. In order to explore the influence factors of internal valve leakage rates, several studies have conducted AE experiments that have assessed the frequency and temporal characteristics of AE signals, valve inlet pressure, valve size, valve type and valve damage [12–14]. However, the above studies only can get the relationship between the multiple factors and leakage rates, but they cannot get the internal valve leakage rates. Kaewwaewnoi et al. did conduct such an experiment and discovered that the root mean square (RMS) in time domain of the leakage signals can be used to predict the valve leakage rate [13]. They established a theoretical model, based on the Lighthill equation, of valve liquids leakage rate. On the basis of this study, Prateepasen et al. established a similar model by observing the relationship between RMS and valve gas leakage rate experimentally [15]. However, in conditions of high noise, low pressure and low leakage rate, these models, which only contain a single parameter, produce inaccurate predictions, such that leakage rates are not evaluated effectively.

In recent years, with the development of artificial intelligence, machine learning algorithms have been used to recognize internal valve leakage rates in a natural gas pipeline. Li et al. [16] and Zhang et al. [17] built intelligent recognition models based on support vector machine for estimating internal valve leakage rates. The models possess the excellent recognition accuracy of 95%. However, long training times are required for these models, which reduce the timeliness of recognition. Accordingly, this study proposes a novel method for recognizing internal valve leakage rates based on AE technology, dimensionality reduction and cluster analysis. It not only can identify the leakage rates, but also improves the recognition speed on the basis of ensuring the recognition accuracy.

Cluster analysis is the process of partitioning a set of physical or

[*] Corresponding author.
E-mail address: zhenlinli@263.net (Z.-L. Li).

https://doi.org/10.1016/j.measurement.2018.04.076
Received 15 September 2017; Received in revised form 20 April 2018; Accepted 20 April 2018
Available online 22 April 2018
0263-2241/

abstract objects into multiple clusters based on similarity. It is also a major data analysis method that is widely used in fields such as target recognition, image classification and information retrieval. It is also used to describe data and measure the similarity between data sources [18–20]. The *k*-means algorithm is one of the classic clustering algorithms [21]. It has the characteristics of easy implementation, simplicity and high efficiency for large data sets, which are the main reasons for its widespread use [22]. In addition, the *k*-means algorithm allows direct parallelisation, and it is insensitive to data sorting [34]. When the resulting clusters are dense and the differences between clusters are distinct, the *k*-means works better than other clustering algorithms. However, the *k*-means algorithm is sensitive to outliers and can easily fall into a local optimum [23]. The *k*-medoids algorithm is an extension of the *k*-means algorithm that does not use the average of the objects in the cluster, but instead selects the centre of the objects in the cluster as the reference point [24]. Due to the central points being selected from the data, the *k*-medoids algorithm is more robust to outliers than *k*-means [25]. *k*-medoids and *k*-means are hard clustering algorithms, where each sample can only belong to one cluster. By contrast, 'fuzzy C means' (FCM) [26,27] is the most widely used fuzzy clustering algorithm. It allows each sample to belong to multiple clusters with a certain degree of sharing. Its natural ability to handle overlapping clusters and the convenience of implementation has seen it succeed in a wide variety of engineering applications [28].

In the above three clustering methods, data samples with high similarity are partitioned into the same cluster. Therefore, the performance of the clustering algorithm is highly dependent on the quality of the extracted features [29]. In this work, eight feature parameters: entropy, energy, skewness, mean, variance, standard deviation (SD), kurtosis and RMS were used as the input variables. High-dimensional features may contain a lot of redundant information, which can reduce computation speed and recognition accuracy. Therefore, in order to improve the performance of clustering algorithms, factor analysis was applied to analyse the correlation of the variables, so as to reduce the dimensionality of the original features and eliminate redundant information [30].

2. Theory

2.1. Acoustic emissions

The *AE phenomenon* is defined as the process of generating transient stress waves through the rapid release of energy from localised deformations or fractures [31]. When valve leakage occurs, a pressure differential at the leakage point causes gas to eject at high-velocity and mix with steady or relatively slow flowing gas, thereby making a loud noise (AE source). AE detection systems have been devised that detect such noises and analyse their features to estimate the valve leakage rates.

2.2. Factor analysis

Factor analysis is a multivariate statistical method based on the dependent relationships between variables. It synthesises variables with complicated relationships into fewer factors so that they are more easily interpreted [30]. Variables are then grouped according to the strengths of the correlations between them. The variables of each group represent a basic structure that is called a *common factor*. The purpose of factor analysis is to look for these common factors, then reduce the amount of redundant information in the highly dimensional data.

Assume we have a dataset $\mathbf{X} = (x_1,x_2,\cdots,x_n)$ that consists of n observable random variables, and a vector $\boldsymbol{F} = (F_1,F_2,\cdots,F_m)$ that consists of m unobserved random variables (common factors). Then the general model of factor analysis can be defined as follows:

$$\mathbf{X} = \mathbf{AF} + \varepsilon \tag{1}$$

where $\boldsymbol{A} = a_{ij}(i = 1,2,\cdots,n;\ j = 1,2,\cdots,m)$ is the loading matrix; a_{ij} is the loading of the ith variable x_i on the jth common factor F_j, which reflects the influence of the common factors on the variables and plays a crucial role in explaining the common factors; and $\boldsymbol{\varepsilon} = (\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$ is a vector that consists of n unobserved stochastic error terms. Eq. (1) should satisfy the following assumptions: (1) $m \leqslant n$, that is, the number of common factors is less than the number of original variables; (2) $Cov(\boldsymbol{F},\varepsilon) = 0$, that is, $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are irrelevant; (3) $E(\boldsymbol{F}) = 0$, $Var(\boldsymbol{F}) = \boldsymbol{I}_{m\times m}$, that is, the common factors are irrelevant to each other and have a unit variance; (4) $Var(\varepsilon) = diag(\delta_1^2,\delta_2^2,\cdots,\delta_n^2)$, $E(\varepsilon) = 0$, that is, the unobserved stochastic error terms are irrelevant to each other.

2.3. K-means

For the dataset $(x_1,x_2,\cdots,x_n)$, each sample is a d-dimensional real vector. The k-means is a centroid-based clustering algorithm. A centroid usually represents the average of the samples contained in a cluster. The algorithm aims to partition the n samples into k clusters $\{C_1,C_2,...,C_k\}$ so as to minimize the objective function. The objective function can be defined as follows:

$$E = \sum_{j=1}^{k}\sum_{i=1}^{n} \|x_i - o_j\|^2 \tag{2}$$

where o_j is the centroid of cluster C_j that is obtained by calculating the average of the samples contained in cluster C_j; and $\|x_i - o_j\|^2$ is the squared Euclidean distance between x_i and o_j, which is mainly used to calculate the dissimilarity between sample data and the centres. Because of this, clustering analysis is sensitive to the selected distance measurement. Therefore, the same clustering algorithm is likely to produce different partition results when different distance measurement methods are used [32].

2.4. K-medoids

The k-medoids clustering algorithm is an iterative, greedy method that partitions the n samples in the sample space into k clusters. It works as follows:

1. k samples are randomly selected as the initial centre points.
2. The remaining samples are associated to the closest centre point.
3. Swap non-centre sample point and centre point. Then compute the total cost of objective function.

Repeat alternating steps 2 and 3 until the objective function is optimal. The objective function can be defined as Eq. (2). Unlike the k-means clustering algorithm, o_j is derived from the sample data. The k-medoids algorithm can handle different types of sample data whilst the clustering results are independent of the order of input samples.

2.5. Fuzzy C means

Fuzzy clustering analysis is one of the main techniques used for unsupervised machine learning. It uses fuzzy theory to analyse and model sample data, thereby establishing the uncertainty of the sample categories and reflecting the real world objectively [33]. It possesses crucial theoretical and practical value, and has been applied to a multitude of fields, such as large-scale data analysis, vector quantization, image segmentation, pattern recognition, etc.

FCM algorithm aims to obtain the membership degree between each data point and all clusters by optimizing the objective function. Then, data point categories are determined so that the data can be partitioned into multiple clusters. The objective function of FCM algorithm can be defined as follows: