2017 International Conference on Modern Education and Information Technology (MEIT 2017)
ISBN: 978-1-60595-468-4

There is a Gold Mine in Flight Data: A Framework of Data Mining in Civil Aviation

Xin-bin ZHAO[1,a], Bin LI[1,b], and Cheng-guo WANG[2,c,*]

[1]China Academy of Civil Aviation Science and Technology, Beijing, China

[2]Yantai Academy, China Agricultural University, Yantai, Shandong, China

[a]zhaoxb@mail.castc.org.cn, [b]lib@mail.castc.org.cn, [c]wangcg@126.com

*Corresponding author

Keywords: Data Mining, Civil Aviation, FOQA, Data Analysis.

Abstract. Preventing unsafe incidents is the focus of civil aviation. Flight data monitoring, also referred to as Flight Operations Quality Assurance (FOQA) is a powerful supporting tool and means for data collection and analysis. FOQA is regularly implemented in civil aviation where the application of data mining methods has been proposed. This paper surveys data mining techniques and aims at the beginner with little foundations of data mining or civil aviation, as well as a review of published work on application to flight data. It then illustrates that flight data is exactly a gold mine, and the application of data mining in FOQA is an opportunity as well as a challenge.

Introduction

Data Mining (DM), popularly known as knowledge discovery in databases, it is the process which can discover the useful, unexpected, valid and understandable knowledge from data. Over the past two decades, it has been rapid strides, especially from the perspective of the computer science community. DM techniques are widely used in many fields, such as the telecommunication industry, economic, financial, healthcare, online retail and marketing, and other scientific applications. In the last decade, the field of civil aviation has achieved rapid development. In China Civil Aviation, for example [1], in 2016, the total number of aircraft increases from 1,000 to nearly 3,000, and the average annual growth rate is 11.3%; the flights of each day are more than 10,000. How to ensure the flights safety is the focus of civil aviation sector.

Onboard Data Recorder (ODR) is the instrument in the aircraft, which is used to record a variety of flight information, such as flight altitude, speed, heading, pitch attitude, flight time and other parameters. The analysis of flight data obtained from ODR can identify the flight status, the pilot operation and mechanical conditions, and it provides support for safety analysis, fault detection, incident examination and efficiency improvement. Flight Data Management (FDM) programs, also referred to as Flight Operational Quality Assurance (FOQA) in numerous countries, have been successfully used for many years. FOQA simply detects the events by some flight parameters which constitute the corresponding event, and concurrently exceeds some prescribed threshold values [2]. The current status of FOQA is facing the following problems and shortcomings: (1) The threshold value of the flight parameters has been made by artificial demarcation which is subjective and could be blind; (2) The identification of the event depends on single flight parameter. For example, it identifies the hard landing event by the single flight parameter—vertical acceleration [3]. When the value of the flight parameter is in the marginal zone, this approach may be largely affected. (3) The uncompleted flight parameter sets of event affects the accuracy of event recognition.

With the development of civil aviation and the advent of the era of big data, human intervention and simple statistical analysis have been unable to meet the demand. Nowadays, DM technology is more and more important. It can assert that the safety benefits of FOQA can be significantly enhanced by DM methods. Though FOQA is taking the lead, and making progresses gradually in the domain of safety analysis, application of DM for safety analysis is still very lack. The emergence of this

phenomenon restricts the development of civil aviation safety analysis, and big data also poses challenges for safety analysis.

This paper shows a brief review of DM. Firstly, it presents the basic concepts and main features of some classical DM methods from different perspective. Then we turn our attention to how to combine DM method with civil aviation safety analysis, and some case studies are provided to illustrate the feasibility of the proposed approach. Lastly, we elaborate on the opportunities and challenges for DM applications in flight data, and emphasize more on the important of the further research.

1. Overview of Data Mining

Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful knowledge from data. Fayyad elaborated the definition of DM in his monograph [4]: "*Data mining […] consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data.*" DM is usually related to computer science, and accomplishes the goal of extracting knowledge through statistical, machine learning, online analysis, information retrieval, expert system and pattern recognition and other methods. A universal DM process, as proposed in [5], is a pipeline containing many phases such as data cleaning and preprocessing, feature extraction, and analytical processing.

1. *Data cleaning and preprocessing:* When the data are collected, they are often not suitable for data processing, to make the data suitable for processing. It is essential to normalize, remove noise, and transform them into a format that is friendly to data mining algorithm, or address missing value.

2. *Feature extraction:* It is crucial to extract relevant feature for the mining process. It can be the greatest degree of eliminating the adverse impacts on the analysis results which caused by error and missing information of the data. Meanwhile DM methods operating directly on high-dimensional space may suffer the so-called curse of dimensionality [6]: handing high-dimensional-samples is computationally expensive and many methods perform poorly in high-dimensional spaces.

3. *Analytical processing:* The final part of the mining process is to design the effective analytical methods from the processed data. Firstly, it needs to identify which DM problem this task belongs to, such as statistical analysis, classification, regression, clustering or correlation analysis; secondly, using the standard DM methods to process the data. But in many cases, it may not be possible to directly use a standard data mining problem. For this case, it is customary to construct a new method which is based on the classical DM methods [7].

DM can be organized in variety ways, as shown in Fig.1. Statistical analysis and knowledge discovery may be distinguished at the idea of data processing. Statistical analysis is typically associated with the traditional statistics, while knowledge discovery can discover the patterns and extract the knowledge from the data, and predict the dynamic of the data. In statistical analysis, descriptive statistics is widely used in FOQA. *Descriptive statistics*' major function is data clearing, i.e. outlier detection. It obtains the data which can reflect the objective phenomenon, and its representative methods are Standard Deviation Criterion (SDC), i.e. Pauta Criterion [8]. Refer to the presence of labels of data (supervised/unsupervised) [9, 10], knowledge discovery technique contains the following four algorithm types: *Classification* is to classify some objects into one of the given categories (classes) [11, 12]. *Regression* can analyze the inherent law of data and build the appropriate dependencies between variables [13, 14]. *Clustering* is aimed at classifying objects into clusters on the basis of their similarity [15, 16]. *Correlation analysis* is to analyze the relationship and extract common feature from two or more variables according to measure the related degree [17, 18].

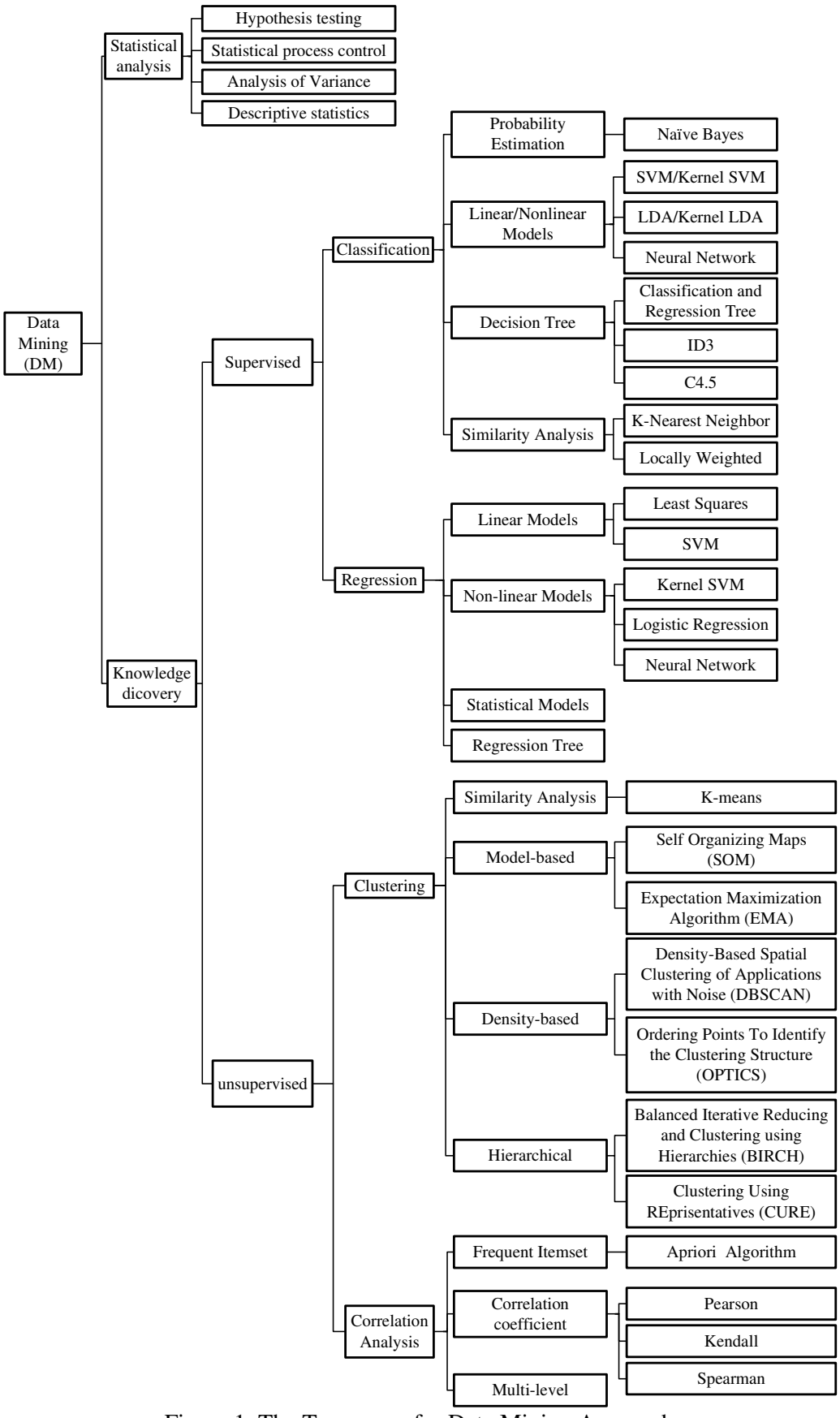

Figure 1. The Taxonomy for Data Mining Approach.

2. Flight Data Mining Applications for FOQA

In this section, the application of descriptive statistics, classification, regression, clustering and correlation analysis in FOQA will be introduced. Each algorithm type consists two parts: algorithm structure and problems corresponding to FOQA.

2.1 Descriptive Statistics---SDC & Quartile Method

The definition of SDC is described as follows: $X \sim N(0,1)$, i.e. random variable X is normal distribution, then

$$P\{-1 \le X \le 1\} = 0.683;\ P\{-2 \le X \le 2\} = 0.954;\ P\{-3 \le X \le 3\} = 0.997. \quad (1)$$

And this concept can be extended to general normal distribution, and the probabilities of X in extents $[\mu-\sigma, \mu+\sigma]$, $[\mu-2\sigma, \mu+2\sigma]$, $[\mu-3\sigma, \mu+3\sigma]$ are 68.3%, 95.4% and 99.7%, respectively. μ is mean which reflects the centralized location of dataset, and σ is standard deviation which shows the range of data distribution [8, 19].

Quartile Method (QM) can be regarded as an improvement of SDC [20], it uses median and Inter Quartile Range (IQR) instead of mean and standard deviation of SDC. Its main statistical parameters includes lower quartile Q_1, upper quartile Q_3, median M, $IQR = Q_3 - Q_1$, upper limit $Q_3 + 3\times IQR$ and sub-upper limit $Q_3 + 1.5\times IQR$. Box-plot is the graph form which can display the collected data.

For FOQA, the threshold value of quality items is crucial. SDC is currently used in this field and has gained wider acceptance. But compared to QM, SDC has plenty of shortcomings (see Table1).

Table 1. Comparison of SDC and QM.

	Outliers Identification	*Distribution*	*Anti-disturbance*	*Visualization*	*Analysis Ability*
SDC	weak	normal	weak	single	weak
QM	strong	none	strong	diversity	strong

In my article [19], we use QM to set the threshold value of quality items, and compare the results of numerical experiment between SDC and QM.

In knowledge discovery techniques, currently, many popular DM methods are designed to use with static data. This so-called static is that data does not change over time or assigns the extreme value of time series to the corresponding parameter [21]. For example, Fig.2 is the flight data of "VRTG" parameter in the "FINAL" flight phase, which is spatially and temporally related. At present, VRTG is represented by one-dimensional value 1.01, which is the maximum value in the data.

In the field of FOQA, to produce more meaningful results, consideration of the spatial and temporal nature of flight data prior to designing DM techniques is paramount [22]. For time series data, there are two relatively easy methods for analysis: raw data analysis and feature extraction analysis. As shown in Fig. 3, raw data analysis can be applied directly on the raw data without modification. In Fig.2, the raw data of VRTG is from 1.00 corresponding 14:12:00 to 0.99 corresponding 14:12:30. Feature extraction analysis extracts the features before using DM techniques, it often as a mean to reduce the large datasets, statistical magnitude is the simple mode in the processing of feature extraction, such as mean, variance, median, standard deviation, coefficient of variation, skewness, kurtosis, maximum, minimum and range.

	FLIGHT_PHASE	VRTG (G)
14:12:00	FINAL	1.00
14:12:01	FINAL	1.00
14:12:02	FINAL	1.00
14:12:03	FINAL	1.00
14:12:04	FINAL	1.01
14:12:05	FINAL	1.00
14:12:06	FINAL	1.00
14:12:07	FINAL	1.00
14:12:08	FINAL	0.99
14:12:09	FINAL	1.00
14:12:10	FINAL	0.99
14:12:11	FINAL	1.00
14:12:12	FINAL	1.00
14:12:13	FINAL	1.00
14:12:14	FINAL	0.99
14:12:15	FINAL	1.00
14:12:16	FINAL	0.99
14:12:17	FINAL	0.99
14:12:18	FINAL	1.00
14:12:19	FINAL	0.99
14:12:20	FINAL	1.00
14:12:21	FINAL	1.00
14:12:22	FINAL	1.00
14:12:23	FINAL	1.01
14:12:24	FINAL	1.00
14:12:25	FINAL	0.99
14:12:26	FINAL	0.99
14:12:27	FINAL	1.00
14:12:28	FINAL	1.01
14:12:29	FINAL	1.01
14:12:30	FINAL	0.99

Figure 2. Flight Data of VRTG.

2.2 Supervised Anomaly Detection---Classification & Regression

Supervised Anomaly Detection (SAD) is a method for anomaly detection where the data contains labels. The data with labels can be used as a training set for decision function which can be attained in both classification and regression.

For classification technique, there are three types of output in FOQA: binary-classification, multi-classification, and probability-classification. Table 2 lists the category which the event belongs to. Classification Tree (CT) [24] and Support Vector Machine (SVM) [11, 21] are the famous methods for classification. CT makes the decision tree to the discrete variable. Its result is a binary or multidimensional tree whose input is a set of training data with label, internal nodes, edges and leaves represent attribute, branch result and class respectively. The goal of SVM is to create a decision function using the training data, and using the decision function to predict which class the testing data belongs to.

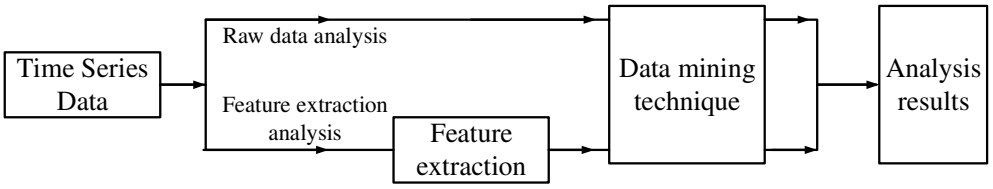

Figure 3. Two Analytical Methods of Time Series Data.

The decision function of Binary-Classification SVM (B-C SVM) is as follows:

$$f(x) = \text{sign}(\sum_i y_i \alpha_i^* K(x_i, x) + b^*) \tag{2}$$

where x_i and x are the training data point and testing data point respectively. α_i is the Lagrangian multiplier for x_i, and b is a biasing term. $K(\cdot,\cdot)$ is the kernel function, the optimization problem can refer to [11]. Multi-Classification SVM (M-C SVM) [11] and Probability-Classification SVM (P-C SVM) [23] can be seen as the extension forms of B-C SVM. The probability output of P-C SVM and the multi-label output of M-C SVM are the main distinction to B-C SVM.

Table 2. The Event Category.

	Binary-classification [11]	Multi-classification [11]	Probability-classification [23]
Event	Runway Excursion Hard Landing Engine Fault Detection	Hard Landing Air Vertical Overload	Tail Strike; Runway Excursion Overload Landing Controlled Flight Into Terrain (CFIT)

Regression is the method which predicts the response variables (dependent variable) using one or more predictive variable (independent variables), that is, according to the linear or nonlinear equation to reflect the relationship between variables. The regression equation has the following form:

$$y = f(x) = \Phi(x) + b \tag{3}$$

$x \in R^m$ and $y \in R^n$ are the predictive variable and the response variable, respectively. $\Phi(\cdot): R^m \to R^n$ is a mapping.

In FOQA programs, the first, regression can be used for feature extraction. For instance, suppose in (3), x represents an event which consists of m parameters, and $y \in R^n$ is a fictitious event with n parameters. When $n < m$, event x achieves feature extraction and data reduction. This process can save the computational time and space in the subsequent data processing. The second, regression can discover of the underlying relations between flight parameters. For x and y are parameters. Regression not only reveals the influence level of x and y, but also can estimate and predict the status of y by x. For example, the form of $\Phi(\cdot)$ in (3) determines the relation of x and y is linear, nonlinear, related strongly or weakly. This application can be used in cause investigation and pilot operation training. The third, regression can reduce noise in data and detect abnormal points using the following equation:

$$d = \frac{\left\| \Phi(x_0) - y_0 + b \right\|_2}{\sqrt{\left\| x_0 \right\|_2^2 + 1}} \tag{4}$$

d is the distance between point (x_0, y_0) and the regression line (3). For empirical values ε_1 and $\varepsilon_2 (\geq \varepsilon_1)$, the class of (x_0, y_0) can be determined by (5):

$$(x_0, y_0) \in \begin{cases} \text{Normal Point}, & d < \varepsilon_1 \\ \text{Abnormal Point}, & \varepsilon_1 \leq d \leq \varepsilon_2 \\ \text{Noise Point}, & \varepsilon_2 < d \end{cases} \tag{5}$$

2.3 Unsupervised Anomaly Detection---Clustering & Correlation Analysis

Cluster aims at classifying objects into clusters on the basis of their similarity, and the objects in the same cluster are similar to each other and dissimilar to other objects in the different cluster. Dataset could generally be divided into two categories: isolated point and clusters (each contains many points). K-means is an unsupervised iterative method [7] which is a classical method in clustering analysis and partitions a given dataset into a user specified number of clusters K. In FOQA, isolated points are either worse or better outliers who represent abnormal or normal events/parameters, respectively. Clusters which contain abnormal points can be recognized by the sample with label. This strategy can improve the efficiency of event analysis, and using this method for data preprocessing is also the direction of big data mining.

Correlation analysis is the method to describe the relationship type and measure the correlation degree of variables. The distinction between correlation analysis and regression is that regression not only reveals the variables' relationship, but also predicts the variables, and correlation analysis is unable to realize prediction. Using correlation analysis techniques, it is easy to get the correlation parameter set of a specific event, and it is very helpful to event investigation, improve flight training quality and deep mining of big flight data. There are two approaches to obtain the correlation parameter set: apriori is a seminal algorithm for finding frequent itemsets using candidate generation [7]; Pearson, Kendall and Spearman algorithms are designed to measure the correlation degree of variables by calculating the correlation coefficient.

Summary

Data mining is a great promise field with a growing domain of applications. The opportunities for data mining techniques transition into FOQA are significant. This paper reviewed the current status of the applications of data mining for flight data, and introduced the future trend of FOQA from the five aspects of descriptive statistics, classification, regression, clustering and correlation analysis. Although flight data analysis based on data mining techniques is not pervasive in FOQA, it is nonetheless receiving increasing attention and is growing. According to the current characteristics of civil aviation, addressing with the adaptation of data mining techniques as well as development of new ones is the research focus. Flight data is a gold mine, and data mining is a tool. Combining data mining and flight data is an opportunity as well as a challenge.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (NO. 11371365, NO. 11301535).

References

[1] Information on http://news.carnoc.com/list/383/383781.html.

[2] Flight Standards Division Civil Aviation Administration of China. The implementation and management flight operational quality assurance (FOQA): AC-121/135-FS-2012-45R1[S].Beijing: Civil Aviation Administration of China, 2015:17-25.

[3] X.B. Zhao, B. Lin, Optimization model of civil aircraft landing vertical acceleration standard, Journal of Shenyang Normal University (Natural Science Edition) 35 (2017) 53-59.

[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, pp. 37-54, 1996.

[5] C.C. Aggarwal, Data Mining: The Textbook, Springer, New York, 2015.

[6] G. Shakhnarovich, B. Moghaddam, Face recognition in subspaces, in S. Z. Li, A. K. Jain (Eds.), Handbook of Face Recognition, Springer-Verlag, New York, 2004, pp. 141–168.

[7] X.D. Wu , V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems 14 (2008) 1-37.

[8] Y. Xue, L.P. Chen, Statistical modeling and R software, Tsinghua University Press, Beijing, 2007, pp. 107-161.

[9] D.C. Tao, X.L. Li, X.D. Wu, W.M. Hu, S.J. Maybank, Supervised tensor learning, Knowledge and Information Systems 13 (2007) 1-42.

[10] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, Proc. Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. II: 264-271.

[11] N.Y. Deng, Y.J. Tian, C.H. Zhang, Support vector machines—optimization based theory, algorithms, and extensions, CRC Press, Boca Raton, 2012.

[12] X.B. Zhao, H.F. Shi, M. Lv, L. Jing, Least squares Twin Support Tensor Machine for Classification, Journal of Information & Computational Science 11:12 (2014) 4175-4189.

[13] S.S. Bu, L. Zhen, J.Y. Tan, X.B. Zhao, G.P. Zhou, A Matrix-based Method for Ordinal Regression, Journal of Information & Computational Science 11:17 (2014) 6209-6220.

[14] X.J. Peng, D. Xu, A local information-based feature-selection algorithm for data regression, Pattern Recognition 46 (2013) 2519-2530.

[15] P. Berkhin, A Survey of Clustering Data Mining Techniques, Grouping Multidimensional Data, (2006) 25-71.

[16] R. Xu, D.Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks, 16 (2005) 645-678.

[17] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonocal Correlation Analysis: An Overview with Application to Learning Methods, Neural Computation, 16 (2004) 2639-2664.

[18] A. Tenenhaus, C. Philippec, V. Frouind, Kernel Generalized Canonical Correlation Analysis, Computational Statistics & Data Analysis 90 (2015) 114-131.

[19] X.B. Zhao, B. Li, Application of outlier detection method in civil aviation early warning: submitted to Journal of Nanjing University of Aeronautics & Astronautics (2016).

[20] C.S. Withers, S. Naciarajah, The distribution and quantiles of the range of a Wiener process. Applied Mathematics and Computation 232 (2014) 766-770.

[21] G.M. Xu, S.G. Huang, Airplane's hard landing diagnosis based on optimized support vector machine, Computer Measurement & Control 19 (2011) 256-259.

[22] J.F. Roddick, M. Spiliopoulou, A survey of temporal knowledge discovery paradigms and methods, IEEE Transaction on Knowledge and Data Engineering, 14 (2002) 750-767.

[23] Q. Tao, G.W. Wu, F.Y. Wang, J. Wang, Posterior probability support vector machines for unbalanced data, IEEE Transactions on Neural Networks, 16 (2005) 1561-1573.

[24] X.C. Wang, X.D. Liu, W. Pedrycz, L.S. Zhang, Fuzzy rule based decision trees, Pattern Recognition, 48 (2015) 50-59.