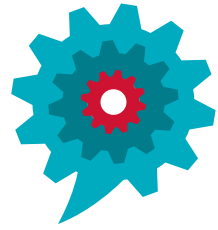
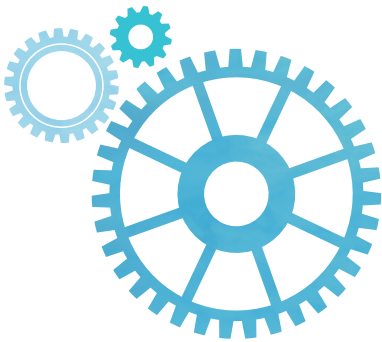


MPMM



Maintenance Performance
Measurement and Management

Proceedings of Maintenance
Performance Measurement
and Management (MPMM)
Conference 2014



4th & 5th SEPT 2014

COIMBRA

Department of Mechanical Engineering
Pólo II · FCTUC



Selection of instances in Condition Based Monitoring

The case of aircraft engines

Leonor Fernandes¹; Roberto Henriques²; Victor Lobo³

¹mlfernandes@euroatlantic.pt; ²roberto@isegi.unl.pt; ³vlobo@isegi.unl.pt

^{1,2,3} ISEGI – NOVA School of Statistics and Information Management of the Universidade Nova de Lisboa
Lisbon, Portugal

³ CINAV- PRT Navy Research Center
Almada Portugal

Abstract— Condition based maintenance (CBM) is based on analysis and data collection monitored by sensors on the aircraft. The knowledge discovery, about the performance of different parameters, by using these data will provide new ways of diagnosing and predicting the state of aircraft engines. However, a single flight produces a huge amount of data that characterize aircraft engine behaviour. The use of algorithms for the simultaneous processing of these data is a difficult and sometimes impossible task. The objective of this work is to choose the best way to select instances for a sample. There should be no loss of relevant information in the sample to identify the state of the engine. We use five methods to select the sample and through clustering techniques and sensibility analysis we choose the best way to select the sample.

Keywords— Condition based maintenance (CBM), aircraft engines, instance selection, knowledge discovery in databases, Self Organizing Maps (SOM).

I. INTRODUCTION

The parameters characterizing aircraft engine behaviour, and operational conditions of the flight, are the types of information that allow changing the usual decision making procedures concerning the removal of engines for maintenance.

Currently, there is much information, gathered over time, about aircraft engines parameters. This information is abundant, so the techniques of Knowledge Discovery in Databases (KDD) are appropriate to describe and characterize the performance of the engines.

But huge amounts of data may be a problem for the use of the algorithms in the data exploratory phase, so different methods of instance selection need to be studied.

This work's objective is to choose the best way to select instances as a sample. We use five methods to select the sample and, through clustering techniques and sensibility analysis, choose the best, i.e. the sample with the smallest dimension, in such a way that the parameters have the same behaviour as the original data.

We use a dataset of engine's parameters for one commercial B767-300 aircraft passenger model, concerning different flights carried out between 2009 and 2013. We also analyze other types of data about the flights' conditions.

The paper is structured as follows: Section II gives a brief description of the maintenance problem; Section III presents a review of KDD. The description of the data and methodology used are in Section IV. Results are presented in Section V. Section VI gives the conclusions of the study.

II. AIRCRAFT ENGINE MAINTENANCE PROBLEM

Maintenance activities of aircraft, particularly the engines, are a possible way to solve aviation companies' trade-off: keeping the aircraft at high security and operability levels versus low costs. The decision to remove an engine to repair usually depends on regular visual inspections, monitoring aircraft engines' parameters and experts' knowledge. The monitoring of an engine's performance parameters is done by comparing the values recorded during the flight with the manufacturer's thresholds. When flight data are close to these thresholds, maintenance actions are taken. Due to the inherent risks of unexpected failure, engine removal is usually done before the optimal time.

One way of modifying this procedure is through extracting more knowledge of engine behaviour parameters and their interactions, as well as the development of designs which are able to predict the future values of these parameters. Thus it is possible to classify the state of an engine and predict the occurrence of the next failure. This engine health management is based on the concept of condition based maintenance (CBM), now a major concern of aviation companies' maintenance departments. [1].

CBM aims, through fault diagnostics and prognostics activities, to assess an engine's current condition and future state. Prognostics deal with tasks that allow the prediction of failure.

Aviation companies currently have a register of a huge amount of data that enables the development of new forms of diagnosis and prognosis of faults. The application of Knowledge Discovery in Databases (KDD) and techniques of Data Mining (DM) to these data of aircraft engines allows the implementation of new CBM policies.

Hundreds of works, theoretical and applied, appear every year in journals and at conferences ([2], [3], [4], [5], [6]) about CBM activities. But the area of diagnosis is the one which is most developed. Prognosis is a relatively new research area [7].

At the end of the last decade three important works about prognostic activity in the industry have appeared [7], [9] and [10]. In the aviation industry, in 2011, [11] prepared a study on operational management in the industry which refers to concerns about prognostic activities.

The followings works shows different approaches to aircraft engines: [8] in 1993 refers to the advantages of the application of neural networks in monitoring an aircraft engine's condition; [9] in 2005 describes a model which would be able as to predict the residual life of an aircraft engine, using the Principal Components Analyses (PCA) technique in oil data; in 2007 two works distinguished themselves – one about the diagnosis of engine failures [10], the other about advanced estimation in engines [11]; in 2009 more studies appeared about the parameters of aircraft engines and their extreme values [12] as well as forms of classifying their faults from the perspective of diagnosis [13]; the following year [14] studied one important aircraft engine parameter, Exhaust Gas Temperature (EGT). They used three different tools: Self Organizing Map, Vector Machine Experts and Genetic Algorithm. In the same year (2010), [15] proposed a new methodology to find the optimal time of replacement safety for critical aircraft parts, using as base a comparison of two models to try to find the best time at which to change important aircraft engine parts; also in 2010 [16] published a work in which they introduced the concept of health management in aircraft engines. They also proposed a new way to predict optimal preventive maintenance interval decision making using a proportional intensity model. Over the last three years a set of works about fault prediction and aircraft engine health monitoring, using Self Organizing Maps, have been presented by [17], [18] and [19].

III. KNOWLEDGE DISCOVERY IN DATABASES (KDD) AND DATAMINING

KDD allows some tasks' performance improvement based on past experience: patterns are identified from historic data. DM includes very important phases: data selection, preprocessing, interpretation and evaluation [20].

During one flight, the amount of available data that allows determining the engine's state is abundant. Different performance and operation parameters of the engines are recorded every second. It is possible to recognize abnormal data in order to identify possible crashes/failures in the engine, but during a flight not all data are relevant.

To select significant data is an important task in the phase of preprocessing of data in KDD [21], but is time-consuming although its tasks contribute to the success of the process [22]. Means for selecting instances are to create a set of training S that belongs to the original data set T. There should be no loss of relevant information in the subset S. S should be able to represent T. The goal is to obtain a sample, with smaller dimensions, in such a way that the variables have the same behavior as the original data [23].

The literature about the instance selection of supervised learning problems is abundant, highlights are provided in [24] which presents a detailed summary of the most commonly used

algorithms with their features and also compares their performances.

However, very little literature (after 2000) was found about instance selection to non-supervised learning. The works of [25], [20], [26] refer to the usual sampling procedures for the selection of the sample (i.e. simple random sampling, sampling uniform, stratified). But instance selection to non-supervised learning, is important. The amount of information, in the data exploratory phase, makes this very analysis complex and ineffective, due to the execution time of some algorithms and the non- implementation of other algorithms [20]. So it is important to develop new ways of instance selection to build the training set S at non-supervised learning.

IV. DATA AND METHODOLOGY

A. Data description

We have 229 flights made by one commercial B767-300 between 2009 and 2013. On each flight were recorded, second by second, the data of 31 performance parameters of the two engines and the flight conditions. The final database is large; there are 4,232,008 instances generating typical DM problems. One of them is instance redundancy in the same flight, so in each flight we want to select only the different instances that are interesting.

Table I describes the parameters of the database, 1/2 are the parameters of engines 1 and 2.

TABLE I – Description of Parameters

Parameters	
<i>Flight Conditions</i>	Altitude
	Gross Weight
	Mach Number
	Static Air Temperature
	Total Air Temperature
	Pressure Total
<i>Engine Performance</i>	Bleed Duct Press Engine 1/2
	Temperature Engine 1/2
	Vibration Engine 1/2
	Engine 1/2 Pressure Ratio
	Request EPR 1/2
	Max Limit EPR1/2
	Fuel Flow Engine 1/2
	Fan Speed-Speed Low Pressure Engine 1/2
	Core Speed-Speed High Pressure Engine 1/2
	Oil Pressure Engine 1/2
	Oil Quantity Engine 1/2
	Oil Temperature Engine 1/2
	Throttle Resolver Angle

The number of observations per flight varies widely because it depends on the duration of each flight. Figure 1 describes the number of instances per flight.

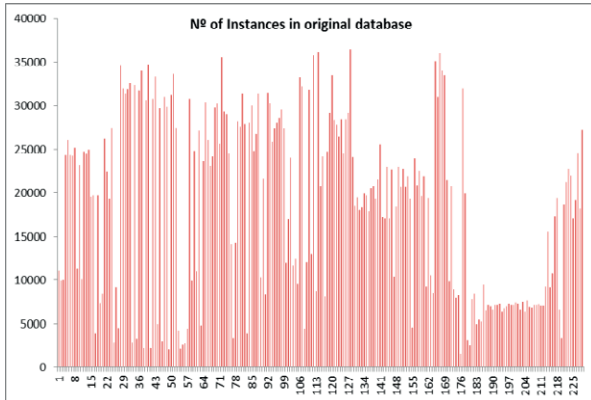


Fig. 1. Distribution of the instances per flight

B. Methodology

We applied five methods to select the sample. Each method is applied by flight. They are:

1. Uniform over time with constant intervals (Systematic Sampling) – original file records, which are ordered per flight, an instance is selected for the sample with a sampling interval of $k = n/N$, where n is the sample size and N the original dataset size.
2. Uniform over time with averages of sets with k instances – systematically select k instances and the value selected for the sample is the average of each of the variables of the k instances.
3. Last Neighbour – the first instance of the original dataset is selected. The Euclidian Distance is calculated between the next instance from the original dataset and the last instance selected to the sample. If this distance exceeds the threshold established, then the instance is selected to the sample and so on.

The threshold has been established through the analysis of the Euclidian Distance between instance i and instance $i-1$ for several flights.

The usual behaviour of the Euclidian Distance on a flight is shown in figure 2. In this flight 17% of the instances have a Euclidian Distance equal to zero.

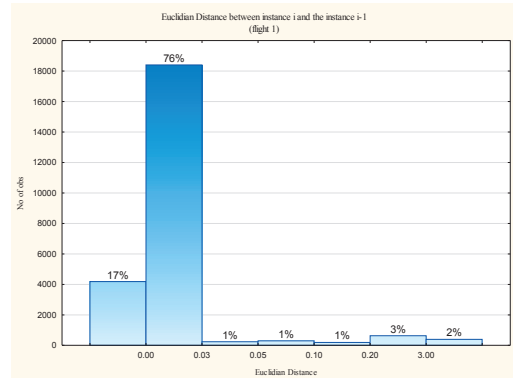


Fig. 2. Distribution of the Euclidian Distance between instance i and instance $i-1$ in one flight.

Table II shows descriptive statistics for three flights.

TABLE II – Descriptive Statistics of the Euclidian Distance between instance i and the instance $i-1$

Flight	Valid N	Mean	Median	Min	Max
1	24329	0.691	0.00	0.00	1213.6
2	26141	0.524	0.00	0.00	1728.5
3	24309	0.608	0.00	0.00	1358.6
Flight	Lower Quartile	Upper Quartile	Percentile 10	Percent ile 90	Coef.Var.
1	0.000	0.001	0.000	0.014	2240.576
2	0.000	0.001	0.000	0.007	2990.828
3	0.000	0.001	0.000	0.006	2680.983

We want instances with big differences so we studied thresholds between 0.03 and 3.

4. All Neighbours – the first instance of the original dataset is selected. The Euclidean’s distances are calculated between the next instance of the original dataset and all the instances selected to the sample. If all these distances exceed the threshold established, then the instance is selected to the sample and so on. The threshold has been established the same way as the previous method.
5. Self-organizing maps (SOM) – Create as many clusters as desired for the sample size. For each cluster compute centroids. These centroids will be the instances of the sample. We process one SOM for each flight.

Several scenarios were made for each method varying their parameters. The analysis undertaken in several flights about the Euclidean Distance between instance i and instance $i-1$, after the data standardisation, is used to establish the starting threshold to methods 3 and 4. It is 0.03. The parameters of the other methods were adjusted in order to obtain samples of nearby dimensions, so that the results can be compared.

The method V is very time-consuming, because that is only applied at last scenario in this work.

To assess the quality of the samples, we use the technique of clustering. In the samples an original database was created for each flight, ten clusters. After this, a Euclidian Distance between the centroids clusters of the samples and the original database was calculated. The total distance of the flight was compared.

We applied the K-means algorithm for the formation of the clusters and chose ten clusters because during the flight there are six natural clusters, the principal phases of flight [25], but in abnormal situations it is possible to have more phases.

V. RESULTS

Each method has different parameters and we simulated several values of these parameters. Table III describes the parameters used and the sample size for the first 4 methods.

TABLE III – Description Methods of Sampling for the first 4 methods

I-Uniform			II-Average Uniform		
parameters	n=sample size	% total	parameters	n=sample size	% total
4	1058002	25	4	1058002	25
5	846492	20	5	846311	20
8	529057	12.5	8	528945	12.5
13	325649	7.7	13	325433	7.7
100	42438	1	100	42215	1
III-Last Neighbour			IV-All Neighbours		
parameters	n=sample size	% total	parameters	n=sample size	% total
0.03	1128388	26.7	0.03	1061178	25.1
0.05	775620	18.3	0.05	735957	17.4
0.1	520821	12.3	0.1	487474	11.5
0.2	327614	7.7	0.2	303666	7.2
3	50851	1.2	3	31127	0.7

Table IV gives the time consumed for each simulation to the first 4 methods.

TABLE IV – Time-consuming, seconds

parameters	Uniform	Average Uniform	parameters	Last Neighbour	All Neighbours
4	17.08	41.60	0.03	711.24	7150.39
5	14.40	34.21	0.05	363.62	3921.23
8	9.88	22.22	0.1	210.39	2153.00
13	7.06	15.05	0.2	181.41	1608.91
100	3.10	5.10	3	201.46	562.63

Low values of the parameters are not of interest because lead to samples of high dimension and are a lot time consuming. So we analyzed the last scenario, which represents

1% of the instances of the original database. For this scenario we used more V method.

Table V shows dimension and the parameters used at V method for the last scenario.

TABLE V – Description Methods of Sampling for the first 4 methods

α	σ	Size of matrix	Number of iterations	Neighborhood Function	n=sample size
0.7-0.02	8-3	229 matrix, the size ranged between [4 5] and [19 20]	20/30	Gaussian	42254

The time consuming by V method was 30264 seconds.

The descriptive statistics of the total distance between the centroids clusters of the samples and the centroids clusters of the original database are in Table VI.

TABLE VI – Descriptive Statistics Total Distance between centroids clusters

Method	N	Mean	Min	Max	Std. Dev.	Coef. Var.
Uniform	229	616.30	13.81	3482.98	704.93	114.38
Average Uniform	229	146.94	13.83	678.32	138.25	94.08
Last Neighbour	229	202.32	68.95	734.49	91.21	45.08
All Neighbours	229	199.23	39.27	989.22	119.14	59.80
SOM	229	76.94	20.794	424.96	75.72	98.41

The Uniform method is the less time consuming but gives the worst values of the descriptive statistics and presents a wide dispersion of total distance. This sample is different from the original database, with the total distance mean being 616.30 because there are some flights with long distances.

The values of average are only representative in the case of Last Neighbour method, where the coefficient of variation is 45.08.

In figure 3 we compare the distance variability per flight in different methods. When all methods are compared we see the predominance of I method and the distance is between 0 and 3500. At second graph cut this method and the distance is between 0 and 1000, there was a significant decrease in the range of the distance. If we compare only the methods carried out with distances between instances the range is the same, and there seems to be no many changes. At last graph we compare only II method and IV method. The distance is less than 700. These two methods have smaller distances, and among them the V method is better.

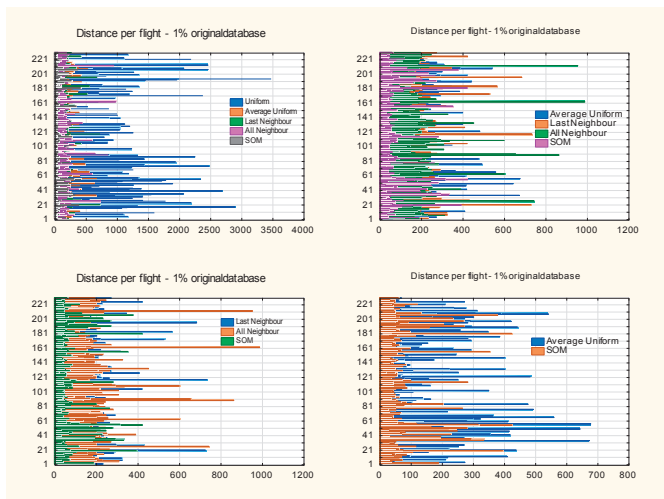


Fig. 3. Comparison of distance variability per flight of the different methods

The figure 4 shows the high dispersion of the total distances of the five methods. The All Neighbour and SOM methods have a low inter-quartile range. The All Neighbour has many extreme values, i.e. this has several flights with the biggest distance between the sample and the original dataset. The SOM, although the extreme values, these don't represent large distances.

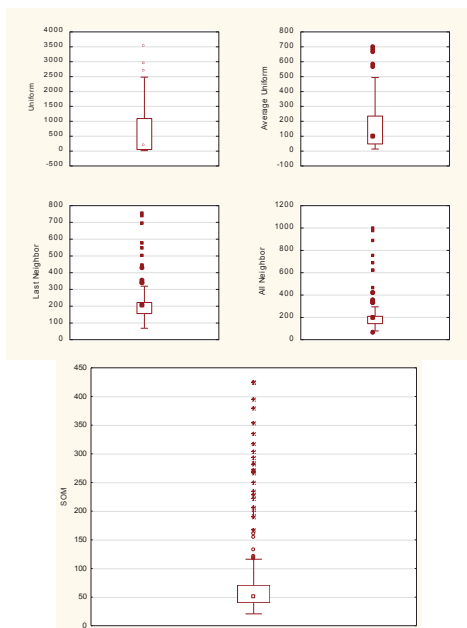


Fig. 4. Distribution of Total distance

The SOM method has the most flights with low distances between the sample and the original dataset as is shows by figure 5.

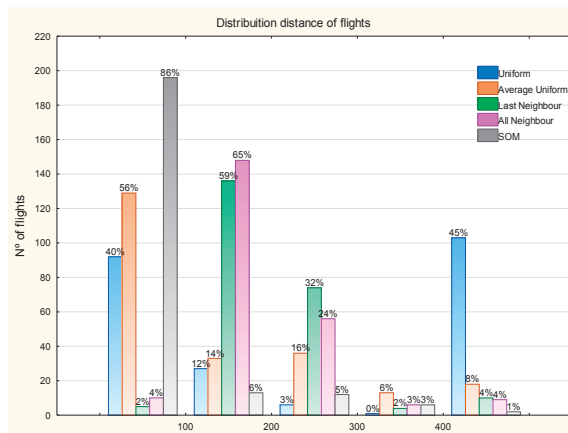


Fig. 5. Distribution distance of flights

The V method is also what has the low total distance (the sum of the distances of all the flights of the sample to the original data set); its distance is half of observed for the II method, figure 6.

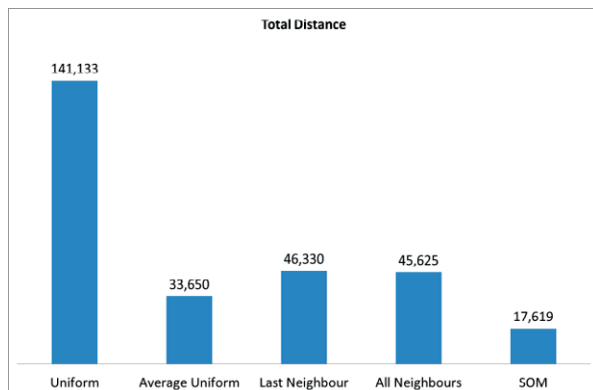


Fig. 6. Total Distance for all flights

VI. CONCLUSIONS

Different methods are used to select the sample close the original dataset but without redundancy instances. The results are analyzed in order of decreasing sample size, time consuming and similarity of original dataset.

In future works, where we intend to apply new procedures to study the engine behaviour parameters and their interactions, we need the sample with these characteristics.

The last three methods used are based on the distance between instances, so they present good values of the dispersion distances. They have a low dispersion for the set of flights.

The method that shows the best performance is the SOM, but it is the most time-consuming. We need use other options for training this neural network to decrease the time consuming.

An alternative method is the Average Uniform. It shows a worse performance than SOM but is the easy implementation and has a good time-consuming.

REFERENCES

- [1] Babbar A.;Syrmos V.L.;Ortiz E.M.;Arita M.M., "Advanced Diagnostics and Prognostics for Engine Health Monitoring," presented at the Aerospace conference, 2009 IEEE, Big Sky, MT 2009.
- [2] Pandian A.; Ali A., "A review of recent trends in machine diagnostics and prognosis algorithms," in *World Congress on Nature & Biologically Inspired Computing*, 2009, pp. 1731-1736.
- [3] Jardine A.K.S.; Lin D.; Banjevic D. (2006) A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*. 1483-1510.
- [4] Thurston M.G., "An open standard for Web-based condition-based maintenance systems," in *AUTOTESTCON Proceedings, 2001. IEEE Systems Readiness Technology Conference*, Valley Forge, 2001, pp. 401-415.
- [5] Ly C.;Tom K.;Byington C. S.;Patrick R.;Vachtsevanos G. J., "Fault diagnosis and failure prognosis for engineering systems: A global perspective," in *Automation Science and Engineering, 2009. CASE 2009. IEEE International Conference on*, 2009, pp. 108-115.
- [6] Kipersztok O.;Dildy G. A., "Evidence-based Bayesian networks approach to airplane maintenance," in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, 2002, pp. 2887-2891.
- [7] Heng A.; Zhang S.;Tan A.;Mathew J., "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mechanical Systems and Signal Processing*, vol. 23, pp. 724-739, 2009.
- [8] Cumming S., "Neural networks for monitoring of engine condition data," *Neural Computing & Applications*, vol. 1, pp. 96-102, 1993.
- [9] Wang W.; Zhang W. (2005, February 11) A Model To predict The residual Life of Aircraft Engines Based Upon Oil Analysis Data. *Wiley Periodicals, Inc.* 276 284.
- [10] Yedavalli R.K., "Robust Estimation and Fault Diagnostics for Aircraft Engines with uncertain model data," in *American Control Confere*, New York, 2007, pp. 2822-2827.
- [11] Viassolo D.E.; Adibhatla S.; Brunell B.J.; Down J.H.; Gibson N.S.; Kumar A.; Mathews H.K.; Holcomb L.D, "Advanced Estimation for Aircraft Engines," in *American Control Conference*, New York, 2007, pp. 2807-2820.
- [12] Babbar A.;Ortiz E.; V. L.Syrmos, "Extreme Value Theory for engine health monitoring and diagnosis," in *Control Applications, (CCA) & Intelligent Control, (ISIC), 2009 IEEE*, 2009, pp. 1069-1074.
- [13] Babbar A.; Ortiz E.M.;Syrmos V.L., "Fuzzy clustering based fault diagnosis for aircraft engine health management,," in *17th Mediterranean Conference on Control and Automation*, Thessaloniki, Greece 2009, pp. 199-204.
- [14] Yukitomo A. R.; Syrmos V. L., "Forecasting gas turbine Exhaust Gas Temperatures using Support Vector Machine Experts and Genetic Algorithm," in *18th Mediterranean Conference on Control & Automation Marocco*, 2010, pp. 345-350.
- [15] Wang P.; Jin T., "Optimal Replacement of Safety-Critical aircraft parts with utilization uncertainties," in *IIE Transaction on Reliability*, 2010.
- [16] Rong X.;Zuo H.;Chen Z., "Civil aero-engine health management integrating with life prediction and maintenance decision-making," in *Prognostics and Health Management Conference, 2010. PHM '10.*, 2010, pp. 1-6.
- [17] Cottrell M.; Gaubert P.; Eloy C.; François D.; Hallaux G.; Lacaille J.; Verleysen M., "Fault prediction in aircraft engines using Self-Organizing Maps," *Advances in Self-Organizing Maps*, pp. 37-44, 2009.
- [18] Côme E.;Cottrell M.;Verleysen M.;Lacaille J., "Aircraft engine health monitoring using self-organizing maps," presented at the Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects, Berlin, Germany, 2010.
- [19] Côme E.;Cottrell M.;Verleysen M.;Lacaille J., "Aircraft engine fleet monitoring using self-organizing maps and edit distance," presented at the Proceedings of the 8th international conference on Advances in self-organizing maps, Espoo, Finland, 2011.
- [20] Liu H.; Motoda H., "On Issues of Instance Selection," *Data Mining and Knowledge Discovery*, vol. 6, pp. 115-130, 2002/04/01 2002.
- [21] Fernández A.; Duarte A.; Hernández R.; Sánchez Á., "GRASP for Instance Selection in Medical Data Sets," in *Advances in Bioinformatics*. vol. 74, M. Rocha, et al., Eds., ed: Springer Berlin Heidelberg, 2010, pp. 53-60.
- [22] Reinartz T., "A Unifying View on Instance Selection," *Data Mining and Knowledge Discovery*, vol. 6, pp. 191-210, 2002/04/01 2002.
- [23] Gama J.; Carvalho A.;Faceli K.; Lorena A.; Oliveira M., *Extração de Conhecimento de Dados* 1ª edição ed. Lisboa, 2012.
- [24] Olvera-López J. A.; Carrasco-Ochoa J. A.; Martínez-Trinidad J. F.; Kittler J., "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, pp. 133-143, 2010/08/01 2010.
- [25] C. Common Taxonomy Team. (2012) Phase Of Flight - Definitions And Usage Notes.