NEW DATA PREPARATION PROCESS – A CASE STUDY FOR AN EXOMARS DRILL

BRUNO RENÉ SANTOS, UNINOVA, PORTUGAL, BRD@UNINOVA.PT
PEDRO TIAGO FONSECA, UNINOVA, PORTUGAL, PTF@UNINOVA.PT
MANUEL BARATA, ISEL, PORTUGAL, MMB@ISEL.IPL.PT
RITA A. RIBEIRO, UNINOVA, PORTUGAL, RAR@UNINOVA.PT
PEDRO A. C. SOUSA, UNINOVA, PORTUGAL, PAS@UNINOVA.PT

ABSTRACT

This paper addresses the data preparation process for a drill fuzzy monitoring tool. The objective is to describe how to automatically generate fuzzy variables, for a drill monitoring system, using inferred nominal values and their dispersion for each variable. Here, we focus on the knowledge discovery tasks that encompass data extraction, data cleaning, data transformation and then the construction of the fuzzy variables.

KEYWORDS: Fuzzy Inference Systems, Data Preparation, Drill

1. INTRODUCTION

ExoMars is the first Aurora [3] program Flagship mission to be assessed by the European Space Agency (ESA). Its aim is to further characterise the biological environment on Mars in preparation for robotic missions and then human exploration. This mission calls for the development of a Mars orbiter, a descent module and a Mars rover [4]. The Mars orbiter will have to be capable of reaching an orbit around the planet Mars. On board will be a Mars rover within a descent module.

The rover will have a drilling system capable of taking soil samples during the drilling process. The aim of our project – MODI – Simulation of Knowledge Enabled Monitoring and Diagnosis Tool for ExoMars Pasteur Payloads – is to develop a monitoring system for faulty behaviours during the drilling process, based on fuzzy logic. The state of the art in drilling systems is focused on the control of the engine behaviour in order to assure that the drill follows previously defined requested *set points*. Our primary objective is to automatically learn the normal drilling behaviour (learning phase) to monitor the drilling process in real time, in order to detect potential faults for alarm generation.

The Fuzzy Logic approach includes a model expressed in terms of a fuzzy vocabulary and where the underlying relationships, between the related fuzzy sets, are represented by rules. Usually, these systems are commonly called fuzzy inference systems (FIS) [5]. The effective model design proceeds from changes in the number, shape and overlapping between fuzzy variables rather than solely from changes in the production rules. In this paper we focus on the knowledge acquisition phase and construction of fuzzy sets of the complete FIS to be developed for the project.

The new data preparation process, which is based on the classical data preparation paradigm [2] is shown in Figure 1. In addition to the classic (crisp) paradigm we added a new task, Fuzzy Set Generation. This last step is needed for the automatic construction of the fuzzy membership functions that represent each variable, which will be used later in the fuzzy rule base system (details about fuzzy set theory and fuzzy inference systems can be seen in [5] and [6].

Hence, the Data preparation steps, proposed for the project, are:

- Extraction: this process is characterized by methods of obtaining data from different databases or files. In our case we use test results from a real drill prototype, which provide the identified variables [1]

- Transformation: data is transformed from a raw state into data most (information) suited for decision support. In our case we removed the headers and unused variables of the data files
- Cleaning: erroneous records are eliminated and data fields are checked for consistency and missing values. In our case we removed all subsets of data that were not important for the nominal value generation, e.g. transition between defined *set points*, initialisations, end of course.
- Integration: data from multiple databases and others sources are integrated into a central data repository. In our case it consists of creating a file with subsets, based on pre-defined scenarios, and calculating the mean and standard deviation for each variable, within the respective scenario.
- Fuzzy sets generation: we use the mean and deviation of each variable within a scenario to construct the respective fuzzy membership function that represents the variable. This approach is based on the nature of the physical process underlying the monitoring objectives: during the drilling process, each variable is expected to present some small random variation in turn of a central value. Variations are due to the environment electrical noise and micro variations of the physical process.

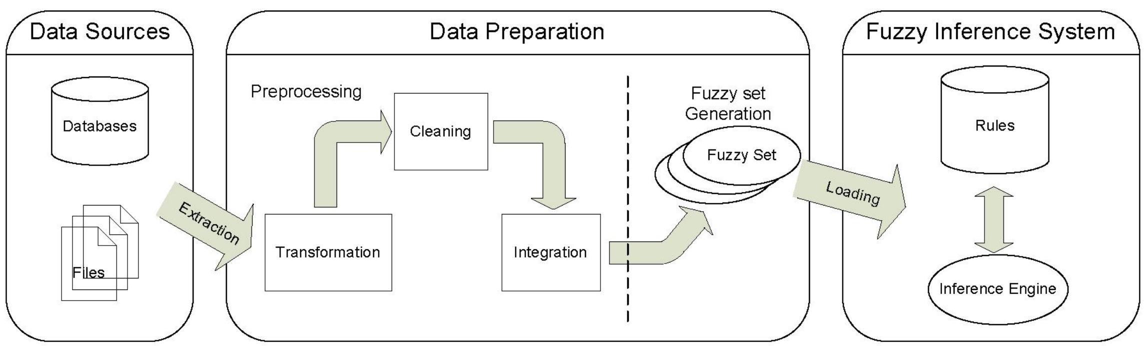


Figure 1 – Data Preparation information flow

Section 2 will present the drill case study. Section 3 presents the steps proposed for the data preparation as well as implementation details of the case study's system. In section 4 we present the conclusions and the planned future work.

2. CASE STUDY

Our subcontractor, Galileo Avionica (Italian company), built a setup for a drill prototype in order to extract data needed for the case study. From this setup it was possible to extract data from eight sensor variables using a synchronized data acquisition system, five different types of materials and a drilling test protocol. The sensor variables are: (1) thrust force; (2 and 3) current for the translational and rotational motor, (4 and 5) voltage for the translational and rotational motor and (6 and 7) speed also for the translational and rotation motor; and (8) depth position.

Each test protocol consisted on combining different speeds for the rotation and translational motors in order to study the behaviour of the system during diverse scenarios and using different kinds of drillable materials. The values used for the speeds will be called *set points*, as they are pre-defined targets, the drill controller will drive the motors in order to achieve this objective during the drilling process. The materials tested were air, gas concrete, concrete, travertine, tuff and marble, i.e. materials enumerated from the softest to the hardest one.

3. DATA PREPARATION PROCESS

On this chapter, all steps concerning data preparation towards nominal value generation will be explained.

3.1. Extraction

The drill setup's sensor acquisition system created data files with synchronized samples extracted from the drilling process on previous enumerated materials. For each one, two types of files were created: (a) extracted data with an information header; and (b) a file with the test protocol followed during the tests.

Regarding the first type of file (a) it has the following characteristics:

- Header: creation date, numbers of samples extracted for each material, sampling rate (on this case study the sampling rate was 10 Hz) and variable names separated by tabs.
- Each line of the file had synchronized values for each variable, separated by tabs.

The test protocol file (b) contains the intervals of time where there was a change on the set points used during the drilling tests. These intervals start with the drill being started until it is turned off. This file also has a header, with general information about the drill. For each relevant time interval, there is the following information: start time, end time and speeds used for the relevant intervals or a written description for the intervals where the drill was being prepared to reach the relevant stages of drilling.

3.2. Transformation

The first transformation on all files consists in removing the header of the file, including the name of the variables. At this stage, the file only has the values for each variable, per column. Any variable that is not relevant for the creation of the fuzzy sets is also removed. As the application used to generate the nominal values was written in Java, there was also the need to change the number separator from comma to a point.

On the protocol test file, the time intervals are represented using minutes and seconds. To create a direct relation between both files, these time intervals are transformed to sample indexes, using Equation (1).

$$sample\_index = sample\_rate \times \left(time(\mathrm{min}) \times 60 + time(\mathrm{sec}) \right) \quad (1)$$

At the time intervals where nothing relevant happens, the set points are initialized to zero. Therefore, the initial protocol test file in raw state, depicted in Figure 2a is transformed into the file depicted in Figure 2b.

```
Trans_Speed nominal = -0,5mm/min
Rot_Speed nominal   = 122rpm

Time (appr.)    Rot_Speed[RPM]  Trans_Speed[mm/min]
---------------------------------------------------
0               Start drilling from new material
04' 00"         122             -0,30
05' 10"         122             -0,48
06' 30"         122             -1,00
07' 40"         160             -1,00
09' 10"         160             -0,42
10' 30"         160             -0,28
12' 00"         Stop drilling
```

a – raw data

```
1       0       0
2400    122     -0.3
3100    122     -0.48
3900    122     -1
4600    160     -1
5500    160     -0.42
6300    160     -0.28
7200    0       0
7213    0       0
```

b- transformed data

Figure 2 – Transformation of data

3.3. Cleaning

Figure 3 shows the evolution of the feed rate during a test within the material concrete. Each recognizable step on the figure relates to a different pair of set points (feed rate, rotation speed) used during a certain time span.

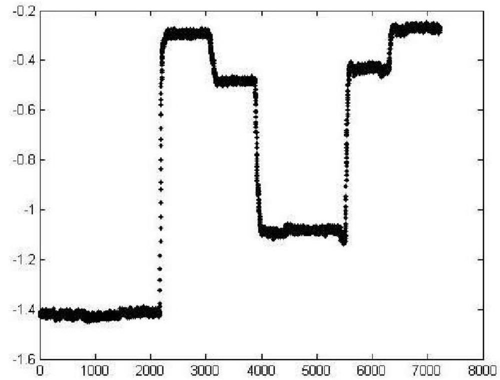


Figure 3 – Feed Rate for Concrete (Translational Speed)

Using the transformed protocol test file, all samples on the data files with intervals where set points are equal to zero, are removed.

3.4. Integration

By parsing the data inside the protocol test files, it is possible to extract the following knowledge:

- Which are the time intervals that contain useful data for the generation of fuzzy sets
- Where each time interval begins and ends

We will call subset to each useful time interval that is relevant to the process at hand (see Figure 3). This process of division also takes place during the data preparation flow of actions.

For each subset we calculate the mean and standard deviation of each variable. Each subset is characterized by having the same set points and material throughout the entire time interval where it was defined. During this time interval, as we already mentioned in chapter 1 and is evidence in Figure 3, the variable values range in a small interval around a mean value. This fact characterizes the behaviour of the monitoring reference we need for the supervision purposes.

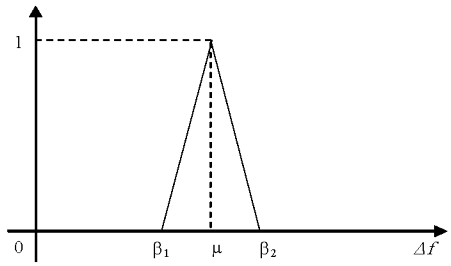
3.5. Fuzzy sets generation

On this section the focus is on the fuzzy sets generation step, where the fuzzy variables are defined. These variables will, subsequently, be used on the fuzzy inference system (FIS), more specifically in the fuzzy rules [5, 6].

For the definition of the fuzzy membership functions that will represent the fuzzification of each variable we use the mean, calculated in task 3.4, for central point. The standard deviation controls the wideness of each membership function.

We consider two types of membership functions. One for the set points (triangular fuzzy sets) and another for the data collected from the sensors (trapezoidal fuzzy sets).

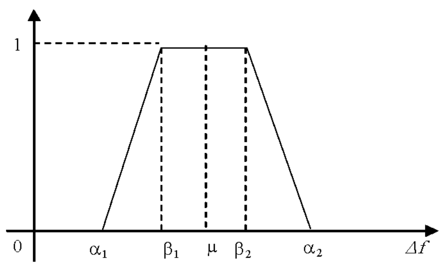
It should be noted that in our case we assume that the fuzzy variables are represented by a single fuzzy set and not by a linguistic variable with several terms [7]. The reason for this is that it is enough to monitor the existence or not of deviations from nominal values and no classification is needed for the deviations.



$$Triangular(\mu) = \begin{cases} \beta 1 = \mu - (\mu \times 10\%) \\ \mu = Nom\_Value \\ \beta 1 = \mu + (\mu \times 10\%) \end{cases}$$

Figure 4 – Triangular Fuzzy set and its creation formula

This fuzzy set representation uses the set point values as its middle point (height of the fuzzy set), which is denoted as μ in Figure 4. The range between β1 and β2 is the allowed range for the set point that a variable can assume, i.e. there is no violation within that range. The values outside from the triangle are considered too big deviations from the pre-defined set points and are considered faulty values.

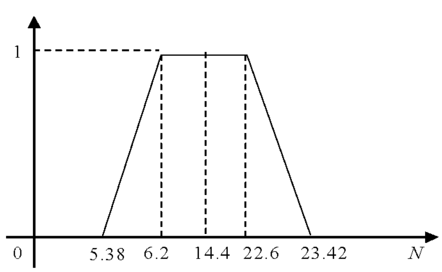


$$Trapzoidal\ (\mu) = \begin{cases} \alpha 1 = \mu - (Calculated\ \ Deviation\ + Calculated\ \ Deviation\ \ *10\%) \\ \beta 1 = \mu - (Calculated\ \ Deviation\ ) \\ \mu = Nom\_Value \\ \beta 2 = \mu + (Calculated\ \ Deviation\ ) \\ \alpha 2 = \mu + (Calculated\ \ Deviation\ + Calculated\ \ Deviation\ \ *10\%) \end{cases}$$

Figure 5 – Trapezoidal Fuzzy set and its creation formula

Besides the set points all other variables are defined as trapezoidal membership functions, defined as in Figure 5. The μ is the nominal value determined in the previous step (3.4) for the variable considered. The range between β1 and β2 is the allowed range for the nominal value that a variable can assume, i.e. there is no violation within that range; the distance between β1&α1 and β2&α2 is the allowed deviation from the nominal value. The further apart from the nominal value, the more penalized the violation will be. The values outside of the trapezoid are considered faulty values.

For example, the fuzzification of the variable "Thrust", which has the nominal value of 14.4 Newtons and a standard deviation of 8.2 (calculated from the previous step), is shown in Figure 6.



$$Trapzoidal(\mu) = \begin{cases} \alpha 1 = 5.38 \quad N \\ \beta 1 = 6.2 \quad N \\ \mu = 14.4 \quad N \\ \beta 2 = 22.6 \quad N \\ \alpha 2 = 23.42 \quad N \end{cases}$$

Figure 6 – fuzzy representation of "Thrust" variable

4. CONCLUSIONS

This paper presents an automatic data preparation approach for the creation of fuzzy sets based on real sensor data. The main advantage of this approach lies on the lack of human interaction, becoming simple to adapt the fuzzy inference system to new problems or scenarios. The approach also showed that, independently of the equipment to be monitored, the fuzzy variables could be determined automatically if there exists test data.

With this approach, the construction of new fuzzy inference systems becomes quicker and less reliant on human expert feedback, which can be difficult to represent with accuracy.

The future work to be done on Data Preparation includes further study on the integration step, with the use of clustering or unsupervised/supervised learning algorithms for the fuzzy sets definition.

5. ACKNOWLEDGEMENTS

This research was developed under the project "Simulation of Knowledge Enabled Monitoring and Diagnosis Tool for ExoMars Pasteur Payloads", contract No: 3-11332/05/NL/GM, of the European Space Agency (ESTEC). We wish to thank our subcontractor Galileo Avionica [1] for the patience on performing all the tests that we require, as well as for the help on identifying the drill variables. We also wish to thank our partner DEIMOS space (Spain) for bearing with us the difficult process of identifying the variables.

6. REFERENCES

[1] G. Avionica, *Homepage*, http://www.galileoavionica.it/, 12-01-2005.

[2] K. J. Cios, W. Pedrycz and R. Awiniarski, *Data mining methods for knowledge discovery*, Kluwer Academic, Boston, 1998.

[3] ESA, *Aurora*, http://www.esa.int/SPECIALS/Aurora/index.html, 12-01-2005.

[4] ESA, *ExoMars*, http://www.esa.int/SPECIALS/Aurora/SEM1NVZKQAD_0.html, 12-01-2005.

[5] J. M. Mendel, *Uncertain rule-based fuzzy logic systems*, Prentice- Hall PTR, 2001.

[6] T. Ross, *Fuzzy Logic with Engineering Applications*, John Wiley & Sons, 2004.

[7] L. A. Zadeh, *The Concept of a Linguistic Variable and its Application to Approximate Reasoning - I. Fuzzy Sets and Applications: Selected Papers by L. A. Zadeh. R. R. Yager, S. Ovchinnikiv, R. Tong and H. T. Nguyen*, John Wiley & Sons, New York, 1987. **Reprinted from: Information Sciences, 8 (1975)**: 219-269