PRELIMINARY ANALYSIS OF LARGE SETS OF PROCESS DATA

Pavel Ettler

COMPUREG Plzeň, s.r.o.

Summary

A decision-support tool for operators of complex industrial processes is being developed in the framework of the long-term international research project. Achievements of the project should help to maintain the best possible production quality under various working conditions. For the project purposes large sets of process data are being gathered which will be used for training and verifying of advanced algorithms based on probabilistic approach.

The contribution deals with preliminary analysis of the data which ought to help to find important relationships among particular signals and to improve understanding of the multi-dimensional problem. Matlab's features such as graphical capability, ability to execute C-coded modules and ActiveX support are being utilized for the task.

1 Introduction

Reliable control system is a must for plants running complex and/or fast industrial processes. Although the well tuned control loops are supposed to ensure flawless production, the final quality of the product still depends on many adjustable parameters and thus on the skills and experience of operators.

The international consortium is attempting to find out correlations between slight quality variations and information hidden in process data. As a result, a decision support tool should be developed to help operators with optimal settings of the process. Academic partners of the project are responsible for theoretical research and development of suitable and powerful algorithms [1] while the application oriented company is committed to implementation [2], [3].

For a pilot application a medium-size cold rolling mill has been selected. For the project purposes a lot of data are being provided for learning and testing of new algorithms. Preliminary analysis of data is a prerequisite.

2 Batch data conversion

Cold rolling on a reversing rolling mill consists in a sequence of passes resulting in reduction of thickness of the metal strip. Data acquisition is triggered synchronously with the strip movement. Every data sample which is available for control is recorded and when a particular pass is finished data are stored in corresponding file having Microsoft Access format.

To allow further processing of data within the Matlab environment a converter from Access to Maltab formats was programmed (Fig. 1). The converter excludes records of sequences which could confuse or devalue further calculations. Because of large volume of data batch processing is used in most cases.

3 Utilization of ActiveX support

For the sake of flexibility an alternative to the batch conversion was sought which would allow to work on data in original format. As a result Matlab's ActiveX support was utilized and a database access control was developed in Visual Basic (Fig. 2). After invoking the control from a Matlab script (or interactively from the command window) it appears in a new figure as shown in Fig. 2. The control utilizes an ordinary common dialog to select an input file. As a result data are transformed into a big CD (Coil Data) matrix within the Matlab's workspace. Number of rows corresponds to number of samples

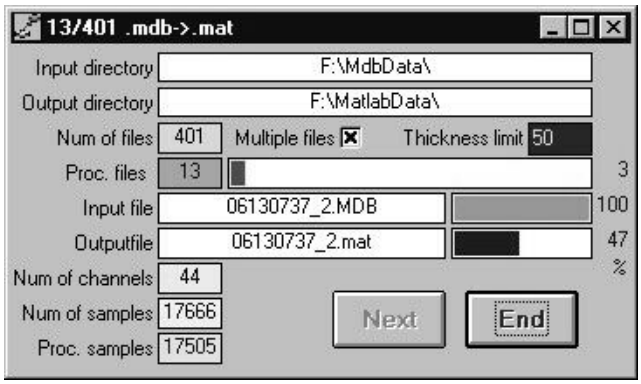

Figure 1: Converter of data files from MS Access to Matlab formats.

while each column contains data from a particular signal channel. Two more columns are created during conversion from encoded logical variables. Auxiliary matrices SignalNames and FileName including signal and original file names respectively are created as well. In the case of selected example the workspace is occupied as follows:

```
Name              Size         Bytes  Class

CD            53406x46      19653408  double array
FileName          1x51           102  char array
SignalNames      46x19          1748  char array

Grand total is 2457601 elements using 19655258 bytes
```

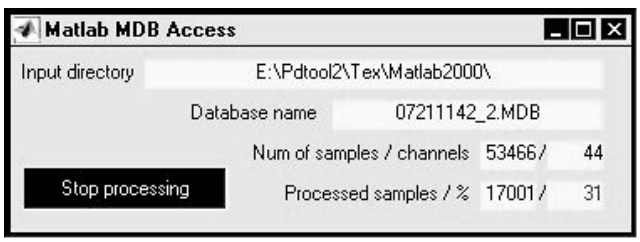

Figure 2: Matlab's figure as a container for database access ActiveX control

4 Quality criteria

Several criteria are being used to evaluate quality of the product. Three statistical coefficients are used mainly:

1. Statistical coefficient C_p defined as:

$$C_p = \frac{tol^+_{h_2} + |tol^-_{h_2}|}{6\ \sigma_{H_2}} ,$$

where H_2 denotes output thickness of the processed metal strip,
h_2 is its deviation from the nominal value H_{2nom},
$tol^+_{h_2}$, $tol^-_{h_2}$ are boundaries of tolerance range of h_2 and
$\bar{H}_2$, σ_{H_2} are mean and standard deviation of the output thickness H_2 respectively.

2. Statistical coefficient C_{pk} defined as:

$$C_{pk} = \frac{\min(\bar{h}_2 - tol^-_{h_2},\ tol^+_{h_2} - \bar{h}_2)}{3\ \sigma_{H_2}} ,$$

where $\bar{h}_2$ denotes mean of h_2.

3. Coefficient C_{per} representing the percentage of the output thickness deviation h_2 being within the tolerance range $\langle tol^-_{h_2},\ tol^+_{h_2} \rangle$.

The aim of the quality control is to maximize coefficients C_p, C_{pk}, C_{per}. For each data file quality criteria have had to be evaluated. For selected data sequences propagation of statistical coefficients has been subjected to investigation as well.

5 Multidimensional clusters and histograms

If the time information is allowed to be eliminated each data sample can be projected in the multidimensional data space where mutually orthogonal axes correspond to single signals. Even if the original number of signals has been reduced to 15 it would be inconvenient to cope with raw data clusters of such dimensions. Transformation into a multidimensional grid of relative frequencies can be used instead. Then each grid cell contains relative number of occurrences of data samples in corresponding subspace given by cell boundaries. Such arrangement can serve as a first rough insight in a conception of multidimensional densities of probability of various working conditions of the machine.

For the sake of reduction of the dimensionality problem all possible 2-dimensional projections were evaluated resulting in 2-D histograms. As a result a set of 15 structures each containing 15 2-D arrays was created. Particular arrays were stored as *sparse* matrices to spare memory (Fig. 4). The workspace is then occupied like this:

```
Name           Size          Bytes  Class

Cp             1x1               8  double array
Cper           1x1               8  double array
Cpk            1x1               8  double array
FileName       1x51            102  char array
h2E            1x1               8  double array
h2nom          1x1               8  double array
h2s            1x1               8  double array
hgSN          15x1            1438  cell array
hg_I           1x1             61320  struct array
hg_I1          1x1             54552  struct array
hg_I1e         1x1             48156  struct array
hg_I2          1x1             57120  struct array
hg_I2e         1x1             52860  struct array
hg_h1          1x1            185304  struct array
hg_h2          1x1            187200  struct array
hg_pz          1x1            110940  struct array
hg_t1          1x1             67824  struct array
hg_t2          1x1             73308  struct array
hg_v           1x1             82212  struct array
hg_v1          1x1            108168  struct array
hg_v2          1x1             78888  struct array
hg_z           1x1             59928  struct array
hg_zs          1x1             65664  struct array
```

```
hg_t1 =
     h1: [100x100 sparse]
     h2: [100x100 sparse]
     v1: [100x100 sparse]
     v2: [100x100 sparse]
      v: [100x100 sparse]
      z: [100x100 sparse]
     I1: [100x100 sparse]
     I2: [100x100 sparse]
      I: [100x100 sparse]
    I1e: [100x100 sparse]
    I2e: [100x100 sparse]
     pz: [100x100 sparse]
     t1: [100x100 sparse]
     t2: [100x100 sparse]
     zs: [100x100 sparse]
```

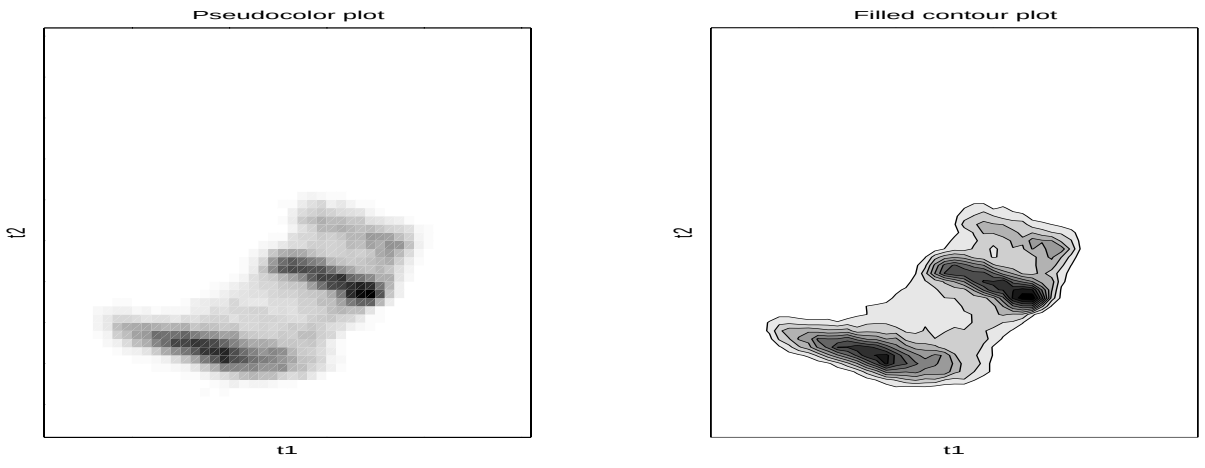

Figure 3: Examples of 2-D histogram visualized as pseudocolor and contourf plots

6 C-coding

Matlab scripts were made to allow automated calculation of

– quality criteria, means, standard deviations and extremes of signals;

– multidimensional grid of relative frequencies distributed into complex structures.

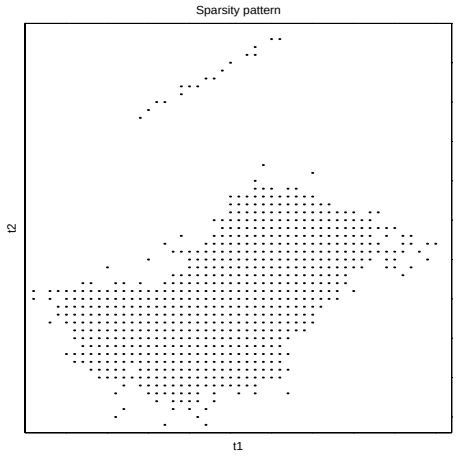

Figure 4: Example of visualization of sparsity pattern

Because many hundreds of files have been waiting for processing the main algorithms were coded in C language and used as mex-functions to speed up calculations. Microsoft compiler and linker were used.

7 Conclusions

Large sets of process data are being analyzed and statistically pre-processed within the Matlab environment. Results will be utilized both for better understanding of the process and for training and verifying of new algorithms which are being developed by other members of an international consortium. Achievements of the project should lead to implementation of the decision-support tool for operators of fast and complex processes such as metal rolling.

8 Acknowledgements

This work has been supported by the grant IST-1999-12058 ProDaCTool of the EC.

References

[1] M. Kárný, I. Nagy, and J. Novovičová: Quasi-bayes approach to multi-model fault detection and isolation. Accepted for publication in: International Journal of Adaptive Control and Signal Processing, 2000.

[2] P. Ettler, F. Jirkovský: Digital controllers for ŠKODA rolling mills. In: Advanced Methods in Adaptive Control for Industrial Applications, Ed.: Lecture Notes in Control and Information Sciences – 158, Springer – Verlag, 1991.

[3] P. Ettler, M. Valečková, M. Kárný, I. Puchr: Towards a Knowledge-Based Control of a Complex Industrial Process. In: Proceedings of the 2000 American Control Conference, Chicago, USA, June 2000.

Pavel Ettler
COMPUREG Plzeň, s.r.o.
Nádražní 18 / P.O.Box 334
306 34 Plzeň
Czech Republic
e-mail: ettler@compureg.cz