

Clustering of an aperiodical medical data

Marek Sacha
sachaml@fel.cvut.cz

Abstract— There were provided aperiodical medical data representing time series of CO₂ levels in blood, information about used drugs and manipulation with a patient after a head harm for analysis. The task was to find possible typical clusters in these time series and also find whether drugs or manipulation influence CO₂ levels. Made experiments using SOM and KMeans algorithms showed distribution of clusters and also importance and influence of drugs and manipulation factors.

I. ASSIGNMENT

The task is to find out typical runs (clusters) of signal CO₂ in provided medical data and also whether drugs or manipulation influence these runs.

II. INTRODUCTION

Medical data provided for semestral work are strictly aperiodical but not random. There are presented data samples of two patients after head harm - particularly development of level of CO₂ in longer time period (several hours) in the datasets.

The task requires to try to find typical time series and influence of drugs and manipulation with the patient. It could be used to allow medical stuff to control the current level of CO₂ more properly.

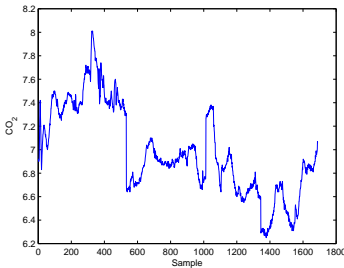


Fig. 1: Time behaviour of CO₂ levels, patient 1.

The datasets comes from diploma thesis¹ focused on neuron sets and time series prediction. The dataset's format is not complicated - each row represents one time-window.

TABLE I: An Example of the dataset

CO ₂ (t-90)	CO ₂ (t-45)	CO ₂ (t-0)	manip(t-0)	drugs(t-25)
6.700000	6.700000	6.780000	6.000000	6.000000
6.650000	6.690000	6.800000	6.000000	6.000000
6.660000	6.680000	6.790000	6.000000	6.000000
6.650000	6.690000	6.790000	6.000000	6.000000
6.660000	6.710000	6.810000	6.000000	6.000000

¹Ing. Josef Bouska: Neuronove site pro predikci casovych rad, https://dip.felk.cvut.cz/browse/pdfcache/bouskj1_2008dipl.pdf

III. METHODOLOGY

There is described a theoretical background of experiments, used tools and tool's configuration in this section.

A. Data mining algorithms

For analysis of datasets, creating and evaluation of models and clustering were used various algorithms. For general knowledge of data structure were used SOM² as there is required *unsupervised learning* for finding relations in time series.

Afterwards there was used KMeans algorithm for finding appropriate clusters according to learned SOM map. Using KMeans it was possible to create a data model and look up other relations in relevant data parts.

B. Tools and configuration

Matlab and SOM toolbox³ were used as software tools. Large scale of functionality of these two tools was used. Datasets were normalized using `som_normalize()` function, size of SOMs was chosen by software automatically according to size of particular dataset. KMeans validation (the best number fo clusters) was also done automatically using *Davies-Boulding index*.

IV. EXPERIMENTS

A. CO₂ levels analysis

First experiment which was made was creating a learned SOM map, using only CO₂ parametrs and it's visualization to get a knowledge of a structure of the data.

It is obvious from the U-matrix that there exist two big clusters (of not properly known inner structure). Top (according to Fig. 2) cluster aproximately represents lower values of CO₂ time-windows where $x \in \langle 4.43, 7.3 \rangle$ and bottom cluster represents higher values where $x \in \langle 7.3, 7.78 \rangle$.

The second step of the experiment was to divide the dataset accroding to these two clusters. For data classification was used KMeans algorithm - searching for up to 5 clusters - low number of cluster prevents overlearning.

The best clustering of 1 to 5 clusters (Fig. 3) according to *Davies-Boulding index* was the same one which was presented in Fig. 2 according to SOM map which is interesting.

The next step in experiment was to discover inner structure of mentioned clusters. Data were separated using matlab into two matrices wich was later used separately again for SOM learning. Results are presented in Fig. 4 and bring some new information.

²Self-organizing map

³<http://www.cis.hut.fi/projects/somtoolbox/>

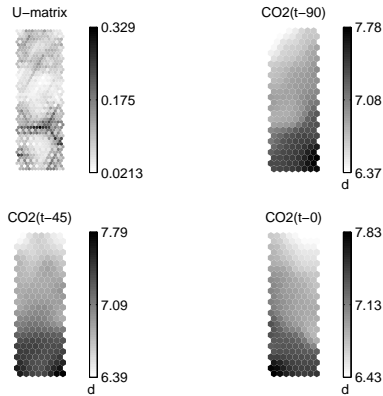
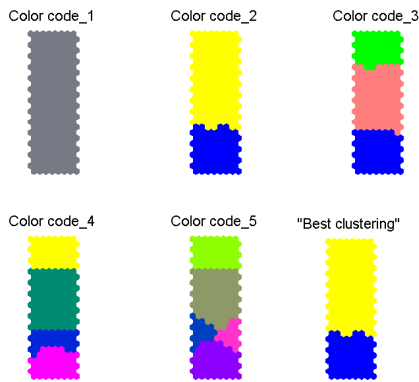
Fig. 2: Learned SOM using CO_2 levels, patient 1.

Fig. 3: Results of KMeans for 1 to 5 clusters, patient 1.

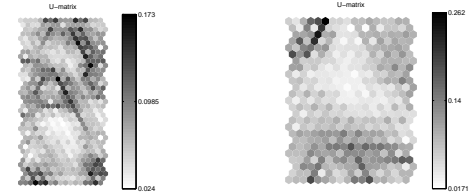
Top cluster ($x \in \langle 4.43, 7.3 \rangle$) has in detail more complicated structure and could be divided into several smaller clusters. On the other hand the majority of the bottom cluster ($x \in \langle 7.3, 7.78 \rangle$) is created by one solid area - this could be also interpreted as similar behaviour (and possibly predictable) of CO_2 levels in a specified range.

B. Influence of drugs and manipulation

The second part of experiments was targeted on finding possible influence of drugs and manipulation to CO_2 levels. It was necessary to prepare another dataset based on input datasets which could help to find appropriate answers. This dataset contained columns *drugs* and *manipulation* and third calculated column *change* - change of level in one time-window: $change = CO_2(t - 90) - CO_2(t - 0)$. This dataset was then also analysed using SOM toolbox.

TABLE II: An Example of a new dataset

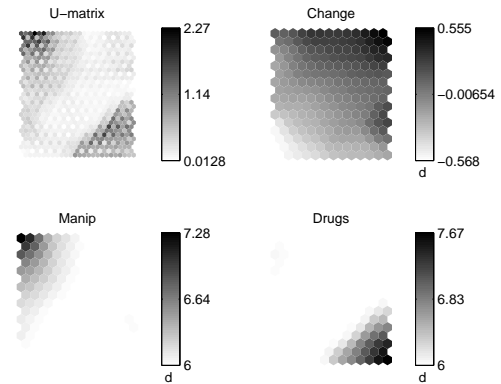
change	manip(t-0)	drugs(t-25)
0.08	6	6
0.17	6.11	6
-0.17	6.219999	6



(a) Top cluster

(b) Bottom cluster

Fig. 4: Inner structure (U-matrices of found clusters)

Fig. 5: Behaviour of CO_2 level according to drugs and manipulation, patient 1.

Visualisation of computed results again brought some interesting information.

1) *Influence of manipulation*: The top-left corner of *change* map visualization is the most important for finding influence of manipulation. It is obvious that influence of manipulation is rather marginal because the values of change in this area vary only around small increase or decrease. This could lead to proclamation that manipulation is not so important variable according to level of CO_2 .

2) *Influence of drugs*: Now is the most important the bottom-right corner of *change* map visualization. It is opposite situation from previous and area could be divided into two parts - darker area representing lower values of drugs and lighter area in the very corner representing high levels of drugs. It is interesting that (according to visualization) lower levels of drugs often led to increase of CO_2 level and on the other hand high levels of drugs could also led to small decrease of the same levels (four neurons in the very corner), but it is definitely obvious that drugs could essentially influence CO_2 levels.

V. DISCUSSION

The proper analysis was targeted only on one of the patients. I did not validate data against the second dataset because of this reason: comparison of only two patients could not discover any common relations, datasets are a lot different and it could be seen from time behaviour plots. From this point of view I decided not to compare because of insufficient amount

of data (datasets for different people) and rather do a proper analysis for one person. The usage of this work I see in the area of further research on similar medical data and inspiration of how to manipulate with the results and eventually prepare more specific datasets to find further relations.

VI. CONCLUSION

The aim of the semestral work mentioned in the introduction was to discover clusters in CO₂ levels time series and to find possible influence of drugs and manipulation to these levels. Using the SOM this work discovered two main clusters in CO₂ level development for lower and higher values and it's simple or complicated inner structure. The second part of the task was fulfilled by creating a new dataset and it's further analysis that brought proof that manipulation is not so important variable but on the other hand drugs could significantly influence both increase and decrease of CO₂ level according to drugs value (type and amount).

REFERENCES

- [1] Ing. Josef Bouska: *Neuronové site pro predikci casovych rad*, FEE CTU, 2008.
- [2] Website of the subject Y336VD
<http://ida.felk.cvut.cz/moodle/course/view.php?id=35>,
FEE CTU, 2008.