ESOM Sampling as a Tool for Detection of Needles in the Haystack of Big Data in Medical Diagnostic Technologies

Alfred Ultsch[1], *Jörg Hoffman*[2] *and Cornelia Brendel*[2]

In particular, within the context of molecular medical research data sets become larger and larger. High data volumes obtained with flow cytometric analyses of blood and tissue samples with real time multiparameter measurements were always a challenge for computer hard and software designers. Today, a regular Flow Cytometry [1] data sat for one single patient typically contains d ($10 < d < 100$) variables for $n > 1,000,000$ single blood cells (counts) [2]. A training period of many years is therefore prerequisite for biologists or physicians who perform the clinical data interpretation. It is, however, clear, that diagnostic structures in these files may be captured by an appropriate sampling procedure. In this work, we compare the advantages and disadvantages for three different sampling strategies producing a dataset consisting of $n_s < 5,000$ as a subset of the n original data: simple random [3], Learning Vector Quantization (LVQ) [4] and a novel proposal based on emergent self-organizing feature maps (ESOM) [5]. For a short overview on sampling strategies, see [3]. The approach is tested on different artificial and experimental datasets. Moreover, we validate our method by performing automated diagnosis of lymphomas employing diagnostic files from original flow cytometric patient lymphoma samples [6].

[1] Data Bionics Research Group, Philipps-University Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany
[2] Dept. of Hematology, Oncology and Immunology, Philipps-University Marburg, Baldingerstrasse; 35043 Marburg, Germany

ultsch@Mathematik.Uni-Marburg.de

References

1 Aghaeepour, Nima, et al. "Critical assessment of automated flow cytometry data analysis techniques." Nature methods 10.3 (2013): 228.

2 Köhnle T., Bücklein, Veit. "Anleitung LAIP Gating Strategie AML" Universities of Munich and Erlangen V 1.2 (2018).

3 Elfil, Mohamed, and Ahmed Negida. "Sampling methods in clinical research; an educational review." Emergency 5.1 (2017).

4 Kohonen, Teuvo. "Improved versions of learning vector quantization." 1990 IJCNN International Joint Conference on Neural Networks. IEEE, 1990.

5 Ultsch, Alfred. "Maps for the visualization of high-dimensional data spaces." Proc. Workshop on Self organizing Maps. 2003.

6 Hoffmann, J. et al "Determination of CD43 and CD200 surface expression can improve diagnostic accuracy of mature B-cell neoplasms" Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaften für Hämatologie und Medizinische Onkologie (2018).