October 2020

Self-Organizing Maps for Quality Control

Anonymous

Self-Organizing Maps for Quality Control

ABSTRACT

This disclosure describes using self-organizing maps (SOMs) to perform unsupervised classification of unlabeled training data with a smaller number of labelled training data. A SOM can then be used to perform quality control of analytical instruments such as mass spectrometers.

KEYWORDS

- Self-Organizing Map
- SOM
- Mass spectrometry
- Mass spectrometer
- Quality control
- Unsupervised learning
- Machine learning

BACKGROUND

A mass spectrometer is an analytical instrument used to measure the mass-to-charge ratios (m/z) of ions of a sample-under-analysis, or an analyte. Typically, the analyte is separated into components via a chromatographic instrument (e.g., via liquid chromatography, gas chromatography, or capillary electrophoresis), the separated components are introduced into an ion source of the mass spectrometer for ionization, and the resulting ions are subject to transport, confinement, and separation by the components of the mass spectrometer for analysis. The analysis can include generating a mass spectrum depicting a plot of intensity (relative abundance) as a function of the m/z. The mass spectrum is useful for the identification, quantification, and structural elucidation of the sample, for example, peptides, proteins, and related molecules.

Over a period of time, the mass spectrometer needs to be cleaned, parts must be replaced, or operating parameters should be adjusted to maintain an expected level of performance. However, it is difficult to determine when maintenance must be performed.

DESCRIPTION

As described herein, instrumentation parameters relating to operation, environmental conditions, and other types of instrument data are used to train a self-organizing map (SOM) as a machine learning technique. Initially, unlabeled instrument data is used to generate the SOM. A small number of labelled instrument data is then used to “seed” the SOM and identify which parts of the classification space defined by the SOM are indicative of an instrument needing maintenance (e.g., cleaning, replacing parts, or adjustment of operating parameters). The SOM is then used by an analytical instrument (e.g., a mass spectrometer) to map its recent instrument data to the classification spaces of the SOM and determine whether maintenance should be performed.

In more detail, a SOM is an artificial neural network (ANN) trained using unsupervised classification to identify undetected patterns in a data set with no pre-existing labels or low (or no) human direction. That is, the SOM is formed by analyzing a data set in which it is not necessarily known whether the instruments that contributed to the data set needed maintenance. This is in contrast with supervised classification schemes in which an ANN is trained using a data set with a known answer or target key (e.g., instrument data that would suggest maintenance should be performed).

For example, an unlabeled data set is first analyzed to generate the SOM. As depicted in Figure 1 below, an N-dimensional sample space representing the unlabeled data set is used to generate a 2-dimensional classification space representing the SOM. One data set that is like another data set would be closer together in the classification space. By contrast, one data set that is very dissimilar to another data set would be farther away in the classification space. Thus, the 2-dimension classification space can have separate regions clustering together different collections of unlabeled data.

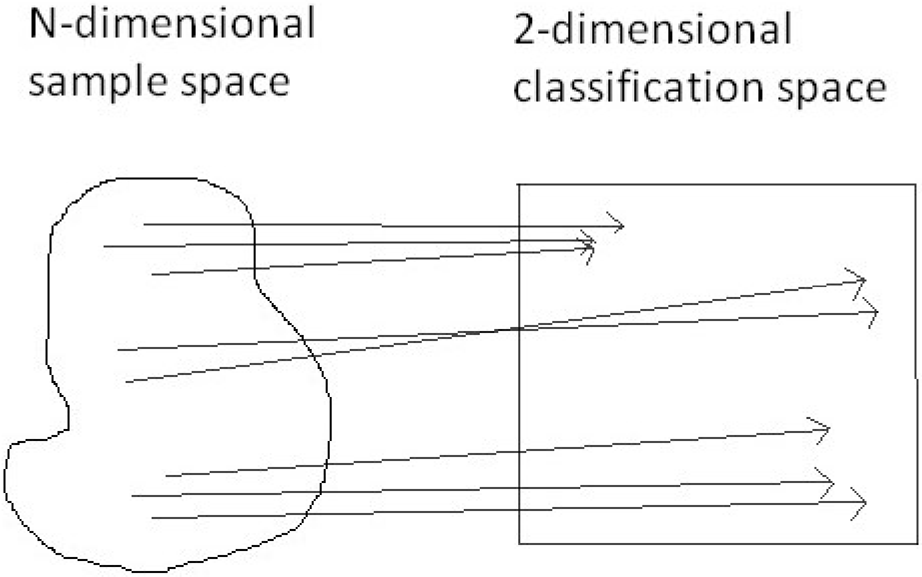
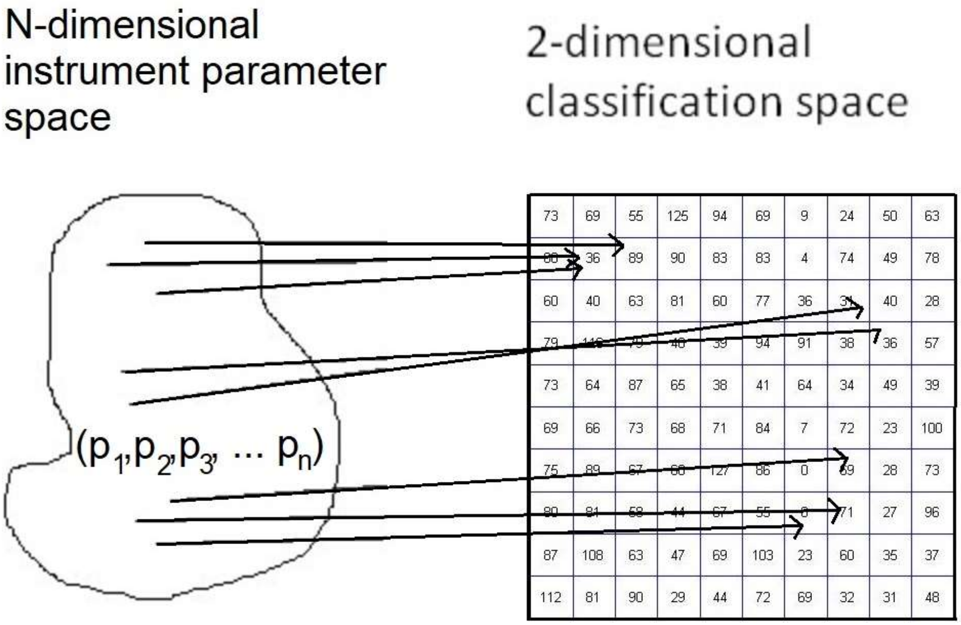

N-dimensional sample space
2-dimensional classification space

Figure 1

In some implementations, the data set includes different categories of data, such as mass spectrometer operation (e.g., related to how MS2, MS3, fragmentation, ionization, etc. were performed), environmental conditions (e.g., temperature, humidity, etc.), type of experiments (e.g., data dependent analysis, data independent analysis, targeted analysis, etc.), type of sample (e.g., biological sample, peptides, lipids, etc.), results of operating (e.g., m/z of molecules, intensities, etc.), or other types of instrument data. Correlations between the data can be used to identify relationships between the data and generate the SOM.

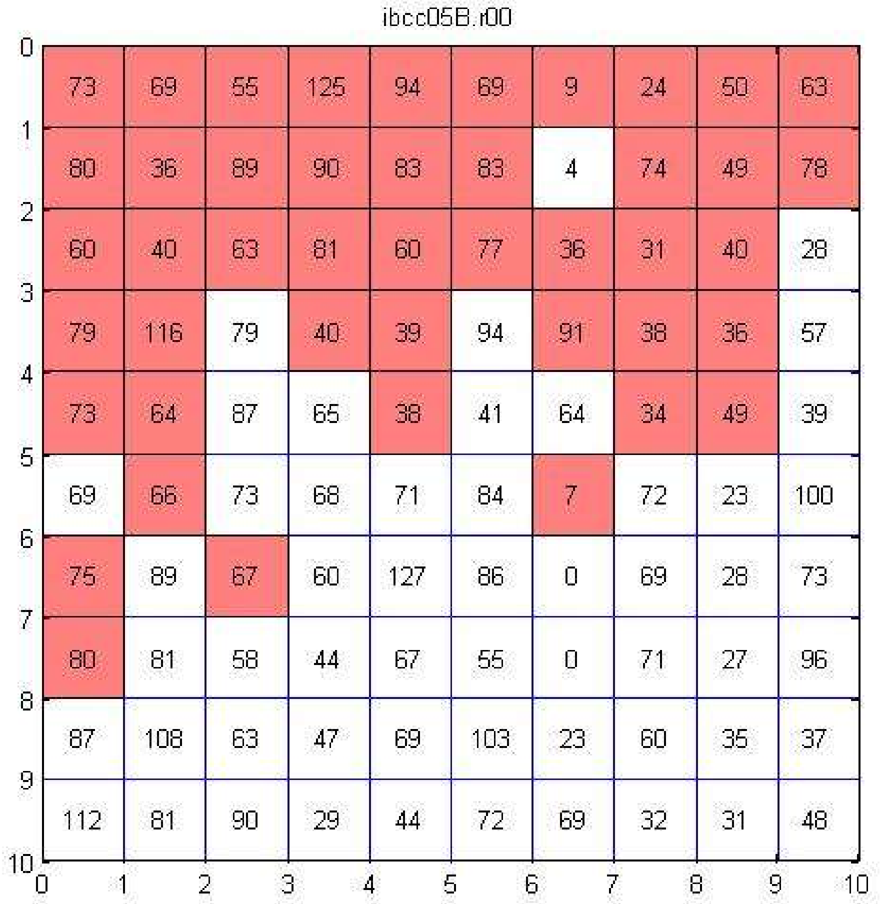
As depicted in example of Figure 2 below, the SOM can be a 10x10 SOM (representing 100 different regions) to represent the unlabeled data set. The numbers in each of the boxes represent how many data sets were mapped to that region of the classification space, with p_1, p_2, p_3,… p_n representing different types of data of a data set. For example, 26 different instrument data might be collected for over a thousand mass spectrometers (or different states at different times of a single or smaller number of mass spectrometers) to provide the unlabeled data set.


N-dimensional instrument parameter space
2-dimensional classification space
$(p_1, p_2, p_3, \ldots p_n)$

73	69	55	125	94	69	9	24	50	63
80	36	89	90	83	83	4	74	49	78
60	40	63	81	60	77	36	31	40	28
79	110	79	46	39	94	91	38	36	57
73	64	87	65	38	41	64	34	49	39
69	66	73	68	71	84	7	72	23	100
75	89	67	60	127	86	0	69	28	73
80	81	58	44	67	55	6	71	27	96
87	108	63	47	69	103	23	60	35	37
112	81	90	29	44	72	69	32	31	48

Figure 2

Next, a labeled data set that includes fewer data sets than the unlabeled data set is used to “seed” the SOM. For example, the labeled cases can be mapped to the appropriate regions within the classification space. Some of the labeled cases might indicate maintenance should be performed, while others might not. Thus, the 2-dimensional classification space can be split in regions that are indicative of maintenance (or even the type of maintenance such as cleaning, replacing a particular part, cleaning a particular part, or adjusting a particular operating parameter such as a voltage applied to an electrode), and some regions that might not be indicative of a mass spectrometer needing maintenance. This is depicted in Figure 3 below, in which the red shaded regions of the classifications pace are indicative of an instrument needing maintenance.



🟥 73	🟥 69	🟥 55	🟥 125	🟥 94	🟥 69	🟥 9	🟥 24	🟥 50	🟥 63
🟥 80	🟥 36	🟥 89	🟥 90	🟥 83	🟥 83	4	🟥 74	🟥 49	🟥 78
🟥 60	🟥 40	🟥 63	🟥 81	🟥 60	🟥 77	🟥 36	🟥 31	🟥 40	28
🟥 79	🟥 116	79	🟥 40	🟥 39	94	🟥 91	🟥 38	🟥 36	57
🟥 73	🟥 64	87	65	🟥 38	41	64	🟥 34	🟥 49	39
69	🟥 66	73	68	71	84	🟥 7	72	23	100
🟥 75	89	🟥 67	60	127	86	0	69	28	73
🟥 80	81	58	44	67	55	0	71	27	96
87	108	63	47	69	103	23	60	35	37
112	81	90	29	44	72	69	32	31	48

Figure 3

The SOM is then used with new instrument data for quality control. For example, the SOM can be provided to a mass spectrometer, and the same types of data used to generate the SOM can be identified for the mass spectrometer as a current data set. The current data set is then mapped to one of the regions of the SOM. If it maps to one of the red shaded regions, then the mass spectrometer might need maintenance.

In some implementations, the new instrument data can be provided to a server which hosts the SOM. The server can determine the quality control and provide the result to the mass spectrometer.

These techniques can be performed on a chromatography or mass spectrometry system, or using a computing system for post-acquisition analysis.

CONCLUSION

Thus, improved quality control can be performed for mass spectrometers. The examples described above involve mass spectrometers, but other analytical instruments such as chromatography systems can also be used. Per the techniques of the disclosure, the SOM data model allows for the quality control to be performed without a significant burden of annotating and labelling a data set.

REFERENCES

1. Kohonen, T., Self-organized formation of topologically correct feature maps, Biological Cybernetics, 43:59-69, 1982