Application of Self-Organizing Maps (SOM) to Chemical Data Analysis

H. Tokutaka*, K. Yoshihara**, K. Fujimura*, K. Iwamoto*, K. Obu-Cann*, T. Watanabe* ,
and S. Kishida*

**Dept. of Electrical and Electronic Engineering, Tottori University,*
Koyama-Minami 4-101, Tottori 680-8552, Japan
Email:tokutaka@ele.tottori -u.ac.jp
**National Research Institute for Metals,1-2-1 Sengen, Tsukuba 305-0047, Japan*
Email:kazuhiro@nrim.go.jp

(Received September 3 1998; accepted January 13 1999)

The SOM method as developed by Kohonen [1] is first applied to problems of chemical analysis. The compositions of the unlabeled spectra whose compositions are unknown can be determined using the SOM method which uses the labeled spectra whose compositions are known. Thus, it has become clear that things that are described qualitatively can be explained more quantitatively. The method can be applied nicely to large input numbers such as a round robin spectra data

KEYWORDS: SOM, data mining, and quantitative chemical data analysis

1. Introduction

Self Organizing Maps (SOM) may be viewed as a lattice or grid with a discrete number of predefined but originally empty sites, referred to as units. During the learning process of SOM, some of the empty units are filled with items. The position of the items on the grid structure depends on the degree of similarity between the items. Similar items are placed on the grid in close proximity. Dissimilar items are placed on units of furthest distance. Each item has multi-dimensional characteristics. The characteristics of different items can be compared by some kind of metric and hence the degree of similarity, i.e. the distance between the items on the grid, can be determined. If an item with incomplete characteristics is placed on the grid, the missing information can be derived from the information that has already been "assigned" (extrapolated) to the empty grid through the learning process.

SOM was developed by Kohonen [1] and is commonly applied to the fields of image and sound recognition. This paper reports on the first studies where SOM is applied to the chemical analysis of alloy compositions. Here, the multi-dimensional characteristics of the items (CoNi alloys) are spectral data from AES (Auger Electron Spectroscopy), as well as the alloy composition. Further, XPS (Xray Photoelectron Spectroscopy) and XRD (X-ray Diffraction) can be considered as multidimensional information sources [2]. The purpose of the SOM method in this application is to derive the alloy composition from the spectral data of this alloy.

2. Quantitative Chemical Data Analysis

2.1 Full Scanned AES Data

Kinetic energy of AES data of Fig.1 from 20 eV to 982 eV by 1 eV division are considered as dimensional units, each spectrum is a 963-dimensional input vector and the normalized signal values between 0 and 1 become signal magnitudes. A SOM similar to that of Fig.2 is constructed using a gray scale. The darker the gray the greater the distance between the nodes. CoNi alloys are grouped from Ni 100% to Co 100%. Cu is next though there is quite some distance between them as shown by the gray level. Ag and Au groups follow respectively.

Conventionally, Auger spectra have been interpreted by subtracting a background and/or separating a main peak from several smaller peaks. The present method considers only the spectral shape as an information source.

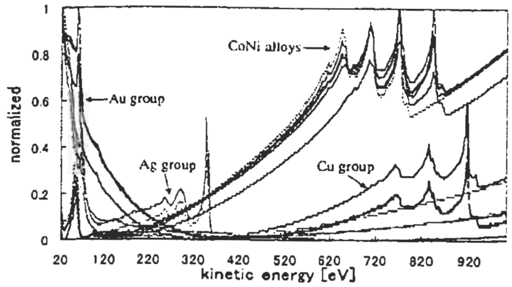

Fig. 1 AES Spectra from metals including CoNi alloys of 6 compositions (Ni 0, 25, 50, 55, 75, 100%).

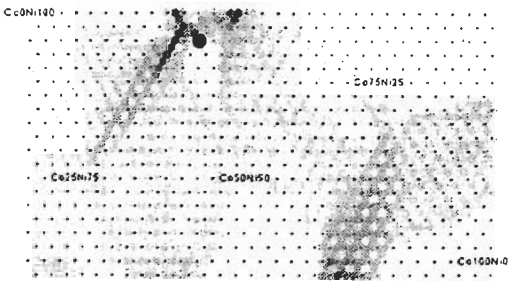

Fig. 2 SOM on 20X30 units using 5 kinds of CoNi Alloys (Ni 0,25, 50, 75, 100%). Large filled circle is unit with best match spectra of Co45Ni55 alloy.

2.2 Characteristic Analysis of High-Tc Superconducting oxides

Surfaces of single-crystal, ceramic and thin films of Bi based high-Tc super-conductors are cleaned by heating and impurities removed. State of cleanliness is assessed by reduction in XPS signal of oxygen (O_2). Binding energies from 520 eV to 540 eV in 0.1 eV increments are considered as 200 dimensions and normalized signal values between 0 and 1 as signal magnitudes. Cleaned surfaces of single crystals heated to 400°C are considered standard clean surfaces and used as standard spectrum. After SOM, three kinds of peaks were identified surrounded by dark gray valleys. These are high-right, split and high-left peaks. Interestingly, the group of high-right peaks consists of heat-treated clean surfaces. Single crystals are the closest group to the standard sample cleaned at 400°C, ceramics follows next. For thin films, heat treated samples at 700 and 800°C are next to ceramics. However, alloys belonging to double peak or high-left peak form groups by their shapes. Here, the oxides are not arranged according to their level of cleanliness, heat treatment or types of oxides. It is concluded that surface impurities affect SOM classifications more than material characteristics.

2.3 Application of SOM to XRD data

SOM was applied to results of XRD from thin films. Samples 1 and 2 are typical experimental data from thin films. XRD results from [2212] (80K phase) and [2223] (110K phase) which are precisely identified single crystal thin films were used as standard. Samples 1 and 2 are not [2212], since the fourth peak (around 22.5°) is not close to these samples. After SOM, [2212] phase is surrounded by a deep gray level, therefore both samples 1 and 2 are not [2212]. However, both samples 1 and 2 are connected with [2223] (110K phase) with a light gray level, although sample 1 is closer to [2223] than sample 2. It is also shown in references [3,4] that more110K phase shows better resistance-temperature (R-T) characteristics.

2.4 Chemical Data Mining

SOM was applied to chemical data mining using AES data of Fig.1. Large backgrounds, which increase in the higher energy region, are subtracted linearly in order to raise the LMM signal sensitivity for CoNi alloys. Energy steps or diffraction angles are considered as dimension units. Known compositions of 6 CoNi alloys in Fig. 1 are considered as new dimensions between 0 and 1. CoNi50% alloy for example, has a dimension of 0.5. Normalized composition values and AES signals ({AES signal –Min AES } which are normalized by the difference between maximum and minimum AES signals) become new signal values. Five samples of Ni

100, 75, 50, 25 and 0 % of CoNi alloys were used as input signals. After SOM learning, all units are compared by the following error function (Err):

$$Err = \sum_{j=1}^{n} (x_j - m_{ij})^2, \qquad (1)$$

where x_j and m_{ij} are the j-th component value of the n-th dimensional input data and i-th unit respectively. Using eq.1, the 5 input data are compared with all the 600 units. All labeled positions in the SOM are determined by the minimum values of eq.1. Co45Ni55 is used as test data where the composition is assumed to be unknown. Using eq.1, all 600 units are compared with Co45Ni55 % spectra data. The unit with the lowest value of eq.1 is identified as the closest unit. For this experiment, the closest unit was 55.25% shown in Fig .2. The actual spectra and learned spectra of Co45Ni55 % are compared in Fig.3, with an error margin of 0.3%.

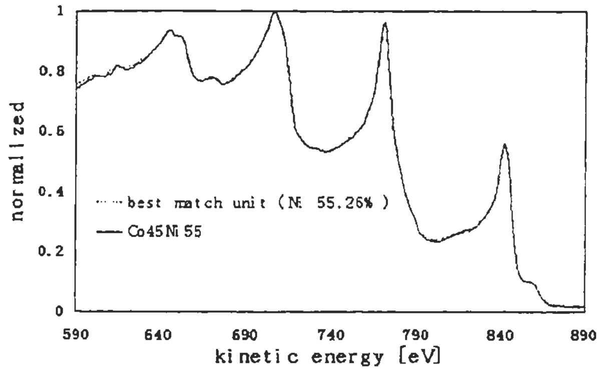

Fig. 3 Comparison between learned unit and original data of Co45Ni55.

3. Higher Number of Item Input

The SOM shown in Fig.2 is constructed with only 5 different items. For the SOM in this section, 90 different items are used. The items are different compositions of CoNi alloys (Ni 0, 25, 50, 55, 75 and 100 %) and from 15 different laboratories. This makes a total of 90 items. The data received from the various laboratories (A, B, C, ... S, X) have different ranges for counts/sec as shown in Fig.4. The data are normalized by the method described in section 2.4 and shown in Fig.5. The SOM based on these 90 items is shown in Fig.6. The SOM of Fig.6 shows a vertical composition separation with gray level expression. Ni 0, 25, 50, 55, 75 and 100 % are arranged vertically separated by gray valleys. In each vertical column, 15 labs. are located from the top to bottom. However, it is only Ni 50% column that has the full number of 15 labs. In other columns, several lab. labels are missing. This is due to the following reason: When several labs. occupy the very same unit in the 600 unit map, the unit is identified using the latest label. It is also very interesting to note that in each vertical column, similar labs., which have similar spectra form a group. The frequency distribution of the composition labels of all 600 units of the map was investigated.

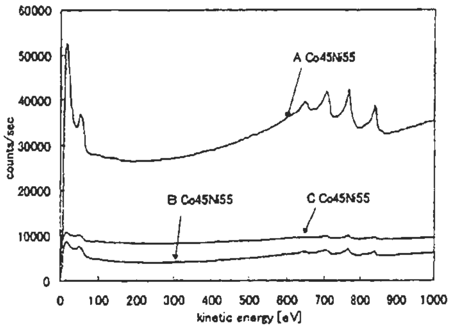

Fig. 4 Input signals of Co45Ni55 from A, B & C Labs.

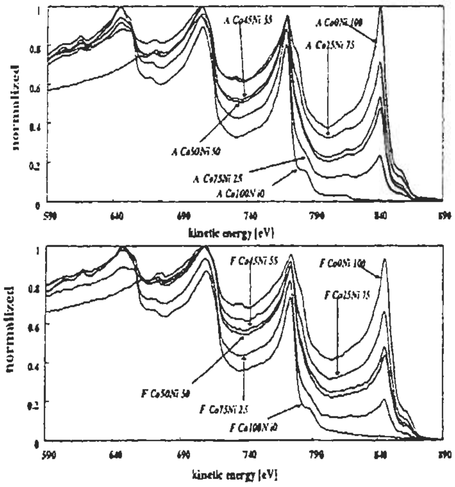

Fig. 5 Input signals of A & F Labs. after background subtraction and normalization within an energy range of 590 - 890eV

The distribution showed that all the compositions were distributed uniformly. Therefore, the SOM map will be suitable for the data mining of any compositions of the alloys. The composition can be determined by finding the best-matched unit.

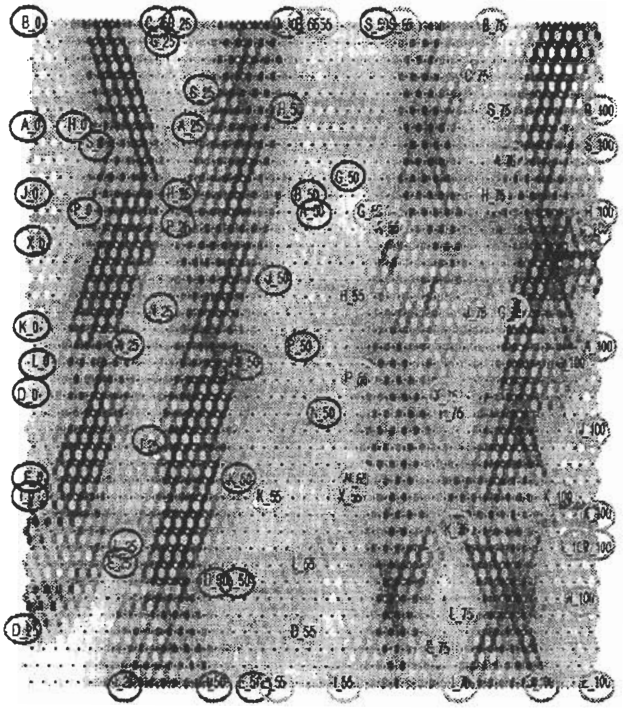

Fig. 6 SOM using 6 different alloys,6 different composition from 15 labs.

4. The Gray Map

The gray map uses gray shade to display the distance of neighboring units. The gray map is a qualitative method only since the gray shade is evaluated merely by the human eye. Now, we propose to use MCP (Modified Counter Propagation System [7], by which all the units on the SOM can be classified to any one of all the input data.. Thus, the discussions are to be extended from SOM to MCP.

5. The Summary

In chemical spectral data analysis, details of spectra and composition analysis can be carried out using methods that consider the physical meaning of the spectra [5]. Analysis can also be discussed by methods such as smoothing by Savitzky-Golay, least square method, multivariate analysis, principal component analysis, etc. [6]. However, SOM is very useful as the first stage of pre-processing before obtaining details using the above mentioned sophisticated methods. The main advantage of SOM is that multi-dimensional input data can be sorted and made visible on 2 dimensional SOM map. The SOM method is applied to determine the chemical composition based on chemical spectra. The method can be tuned to determine the composition and the spectra for the units using the input spectra. If the MCP is used as a supplementary method for the gray level map, the neighboring relation becomes clearer and simpler. Thus, the introduction of SOM to chemical analysis allows us to explore a new development from which we expect future progress.

6. References

[1] T. Kohonen, Self-Organizing Maps (Springer Series in Information Sciences, Vol.30,1995).

[2] H. Tokutaka, Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems (ICONIP'97), 1318 (1997).

[3] A. Shibasaki, S. Kishida, H. Tokutaka, K. Fujimura, and H Naoe, Proceedings of the 9th International Symposium on Superconductivity (ISS'96), 403 (1996).

[4] K. Harada, S. Kishida, T. Matsuoka, T. Maruyama, H. Tokutaka, and K. Fujimura, Jpn. J. Appl. Phys. 35 Part 1, 4297 (1996).

[5] D. Briggs and M. P. Seah, Practical Surface Analysis (John Willey & Sons, 1983).

[6] K. Sasaki, S. Kawata, and S. Minami, Appl. Opt. l.23, 1955 (1984).

[7] R. Hecht-Nielsen, Neurocomputing (Addison -Wesley Publishing Co., Inc., 1993).