Visualization of Historical Process Data of Polyethylene Production by Regularized Fuzzy c-means Clustering

Cs. Vincze, J. Abonyi*, S. Migály, P. Árva, and S. Németh

Department of Process Engineering, University of Veszprém,
P.O.Box 158., Veszprém 8200, Hungary,
*abonyij@fmt.vein.hu www.fmt.vein.hu/softcomp

Abstract
This paper presents a new fuzzy clustering algorithm for the clustering and visualization of high-dimensional data. The cluster centers are arranged on a grid defined on a small dimensional space that can be easily visualized. The smoothness of this mapping is achieved by adding a penalization term to the fuzzy c-means (FCM) functional. Coding the values of prototypes with interpolated black and white colors, regions with different colors evolve on the map and the relation between the variables reveal. The proposed approach is applied in the analysis of an industrial polyethylene plant. The results show that the proposed algorithm is able to detect the relations between the process variables and the quality of the manufactured polymer.

1 Introduction

Clustering based computational intelligence methods are becoming increasingly popular in the pattern recognition community. They are able to learn the mapping of functions and systems, and can perform classification from labeled training data as well as explore structures and classes in unlabeled data. The visualization of high-dimensional data is also important pattern recognition task. Advanced visualization tools should be able to convert complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display and compress information while preserving the most important topological and metric relationships of the primary data items.

Among the wide range of possible tools, the self-organizing map (SOM) is one of the most effective [1]. The Self-Organizing Map as a special clustering tool provides a compact representation of the data distribution, has been widely applied in the visualization of high-dimensional data. SOM implements an ordered mapping of the high-dimensional distribution of the data onto a low-dimensional grid. The SOM algorithm can be considered as a generalized version of the hard c-means clustering algorithm.

Hard clustering methods are based on classical set theory, and it requires an object that either does or does not belong to a cluster. Fuzzy clustering methods operate with fuzzy sets allowing the objects to belong several clusters simultaneously with different degrees of membership [2]. The data set is thus partitioned into c fuzzy subsets. In many real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several classes are not forced to fully belong to one of the classes. Recently, several approaches have been worked out to increase the performance of SOM by the incorporation of fuzzy logic. In a study by Vuorimaa [3], replacing the neurons with fuzzy rules, allowing an efficient modeling of continuous valued functions modified the SOM algorithm. In [4] fuzzy clustering combined with SOM is used to project the data to lower dimensions. Chen-Kuo Tsao et al. [5] integrate some aspects of the fuzzy c-means model into the classical SOM framework. Finally, in [6], a fuzzy self-organizing map is presented based on the modifications of the fuzzy c-means functional. In this approach, the code vectors are distributed on a regular low-dimensional grid, as in SOM and a penalization term is added in order to guarantee a smooth distribution for the values of the code vectors on the grid. The idea of the ordering of the clusters in a smaller dimensional space can be also found in [7], where the fuzzy c-means functional has been modified to detect smooth lines.

The aim of this paper is to generalize the idea of smoothly distributed fuzzy clustering [6] and the Fuzzy curve trace algorithm (FCT) [7] and to apply it to the visualization of high-dimensional process data.

2 Regularized Fuzzy c-means Algorithm

In this paper the clustering of quantitative data is considered. The data are typically observations of some physical phenomenon. Each observation consists of n measured variables, grouped into an n-dimensional column vector $\mathbf{z}_k = [z_{1k},\ldots,z_{nk}]^T, \mathbf{z}_k \in \Re^n$. A set of N observations is denoted by $\mathbf{Z} = \{\mathbf{z}_k | k = 1,2,\ldots,N\}$ and represented as a $n \times N$ matrix. In the pattern recognition terminology, the columns of Z called patterns or objects, the rows are called the features or attributes, and Z is called the pattern matrix.

The objective of clustering is to divide the data set Z into c clusters. A $c \times N$ matrix $\mathbf{U} = [\mu_{ik}]$ represents the fuzzy partitions where $\mu_{ik} = [0,1]$ denotes the degree of the membership of the $\mathbf{z}_k = [z_{1k},...,z_{nk}]^T$ -th observation belongs to the $1 \leq i \leq c$ -th cluster.

The objective of fuzzy c-means clustering is to minimize the sum of the weighted squared distances between the data points, $\boldsymbol{z}_{\mathrm{k}}$ and the cluster centers, $\boldsymbol{v}_{\mathrm{i},}$, where the distances $D_{i,k}^2$ are weighted with the membership values $\mu_{i,k}$, where $D_{i,k}^2$ can be determined by any appropriate norm, e.g., an A-norm:

$$D_{ik}^2 = \|\mathbf{z}_k - \mathbf{v}_i\|_{\mathbf{A}} = \sqrt{(\mathbf{z}_k - \mathbf{v}_i)^T \mathbf{A} \ (\mathbf{z}_k - \mathbf{v}_i)} \qquad (1)$$

Usually spherical clusters are applied when A is an identity matrix, A=I.

Based on the prebious considerations, the objective function of the fuzzy c-means algorithm is

$$J(\mathbf{Z},\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{i,k}^2 \qquad (2)$$

Where $\mathbf{U} = [\mu_{ik}]$ is a fuzzy partition matrix of Z, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2,..., \mathbf{v}_c]$ is a matrix consists of the cluster prototype vectors (centers), and $m \in \langle 1,\infty)$ is a weighting exponent that determines the fuzziness of the resulting clusters and it is often chosen as *m=2*.

The minimization of the c-means functional (Eq. 2) with respect to the following constraints

$$0 < \sum_{k=1}^{N} \mu_{ik} < N, \ 1 \leq i \leq c \sum_{i=1}^{c} \mu_{ik} = 1, \ 1 \leq k \leq N \qquad (3)$$

Represents a non-linear optimization problem that can be solved by using a variety of available methods [2]. The most popular method, however, is the alternating optimization (AO), known as the fuzzy c-means algorithm (FCM-AO):

The fuzzy c-means clustering algorithm is able to detect groups in the data, but the obtained clusters are not ordered which makes the interpretation of the model to be difficult. The aim of this paper is to increase the transparency of the result of clustering by ordering the cluster centers (prototypes) on an easily visualizable low (in this paper two) dimensional space.

According to this motivation, similarly to SOM, the cluster centers (code vectors) are distributed on a two-dimensional lattice, but other topologies can be considered. The proposed algorithm performs a topology preserving mapping from high, *n,* dimensional space of **z**, onto the small, $s<n$ dimensional map, **x**, of the cluster centers such that the relative distances between the data points are preserved. The cluster centers of the map are connected to the adjacent cluster centers by a neighborhood relation, which dictates the topology of the map. For instance, Figure 1. shows a regular square grid, corresponding to $c=9$ code vectors.

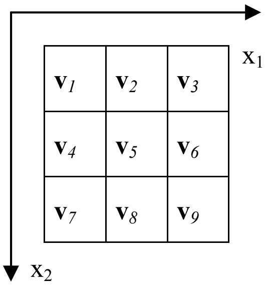

Figure 1. Example of cluster centers arranged on a two-dimensional grid

The proposed arrangement of the clusters makes the resulted model interpretable only if the smoothness of distribution of the code vectors on the grid of the smaller dimensional space is guaranteed. This ordering can be achieved by the regularization of the clusters [13]. To achieve such regularization a penalization term should be added to the original FCM objective function.

$$J(\mathbf{Z},\mathbf{U},\mathbf{V})=\sum_{i=1}^{c}\sum_{k=1}^{N}\left(\mu_{ik}\right)^{m}D_{i,k}^{2}+\vartheta\sum_{i=1}^{c}\left(\left\|\frac{\partial^{2}\mathbf{v}_{i}}{\partial\mathbf{x}^{2}}\right\|\right) \qquad (4)$$

where the smoothness is measured as the second-order derivative of the cluster centers, $S=\int\left\|\frac{\partial^{2}\mathbf{v}}{\partial\mathbf{x}^{2}}\right\|d\mathbf{x}$ and $\vartheta>0$ is the regularization parameter.

Since the additional regularization term can be approximated by the second order differentiation, this part of the cost function can be written in matrix multiplication form:

$$J(\mathbf{Z,U,V})=\sum_{i=1}^{c}\sum_{k=1}^{N}\left(\mu_{ik}\right)^{m}D_{i,k}^{2}+\vartheta\mathbf{VLV}^{T} \qquad (5)$$

where **L** is computed as:

$$\mathbf{L} = \mathbf{G}^T\mathbf{G} \quad \mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \vdots \\ \mathbf{G}_\mathrm{s} \end{pmatrix} \tag{6}$$

Where $\mathbf{G}_\mathrm{i}$ denotes the second order partial difference operator in the direction of the x_i, $i = 1,...,s$ variable with some boundary conditions. At the boundary the second order difference doesn't exist that's why there are so many zeros in the matrices (7). In case of the model depicted in Figure 1., these matrices are the following:

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{G}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

For fixed **U,** the minimum of the cost function with respect to $\mathbf{v}_\mathrm{i}$ $i = 1,...,c$ is the following system of linear equations:

$$\mathbf{v}_i \sum_{k=1}^{N} \mu_{ik}^m + \vartheta \sum_{j=1}^{c} L_{ij} \mathbf{v}_j = \sum_{k=1}^{N} \mu_{ik}^m \mathbf{z}_k \tag{8}$$

where L_{ij} denotes the *i,j*-th element of the **L** matrix.

Represents a non-linear optimization problem that can be solved by using a variety of available methods [2]. The most popular method, however, is the alternating optimization (AO), known as the fuzzy c-means algorithm (FCM-AO). Based on the previous equations, the algorithm of the modified clustering algorithm is given in Table 1.

Table 1. Regularized fuzzy c-means algorithm

Initialization:

Given the data set **Z**, choose the number of clusters *c*, the weighting exponent *m,* the termination tolerance $\varepsilon > 0$.

Repeat for $l = 1,2,...$

Step 1. Compute the cluster centers:

$$\mathbf{v}_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m \mathbf{z}_k - \vartheta \sum_{j=1, j\neq i}^{c} L_{ij} \mathbf{v}_j}{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m + \vartheta L_{ii}}$$

Step 2. Compute the distances:

$$D_{ik}^2 = \left\| \mathbf{z}_k - \mathbf{v}_i \right\|^2 = \sqrt{\left(\mathbf{z}_k - \mathbf{v}_i\right)^T A\left(\mathbf{z}_k - \mathbf{v}_i\right)},\ 1 \le i \le c,\ 1 \le k \le N$$

Step 3. Update the partition matrix:

If $D_{ik} > 0$ for $1 \le i \le c,\ 1 \le k \le N,$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}} \qquad \text{Otherwise } \mu_{ik}^{(l)} = 0$$

The clustering algorithm is sensitive to variations in the numerical ranges of different features. Therefore, the clustering was performed based on normalized data, where all transformed features have zero mean and unit variance, $\tilde{z}_{j,k} = \frac{z_{j,k} - \bar{z}_j}{\sigma_j}$, where $\bar{z}_j$ and σ_j are the mean and the variance of the j-th feature.

3 Classification of polymer products

The data comes from a continuous polyethylene manufacturing plant at the TVK. Ltd. The data is labeled according to the different product grades of a Phillips Petroleum Co. suspension ethylene polymerization process. The polymer particles are suspended in an inert hydrocarbon. The catalyst and the inert solvent are introduced into the loop reactor where ethylene and an α-olefin (hexene) are circulating. The inert solvent (isobuthane) is used to dissipate heat, as the reaction is highly exothermic. A cooling jacket is also used to dissipate the heat of the polymerization. The main properties of polymer products (Melt Index (MI) and density) are controlled by the reactor temperature, monomer, comonomer and chain-transfer agent concentration.

An interesting problem with the process is that it is required to produce different product grades according to market demand. Hence, there is a clear need to minimize the time of changeover because off-specification product may be produced during transition. The difficulty of the problem comes from the fact that there are more than ten process variables to consider. Measurements are available in every 15 seconds on process variables, which are the reactor temperature (T), ethylene concentration in the loop reactor (C2), hexene concentration (C6), the ratio of the hexene and ethylene inlet flowrate (C6/C2in), the flowrate of the isobuthane solvent (C4), the hydrogen concentration (H2), the density of the slurry in the reactor (roz), polymer production intensity (PE), and the flowrate of the catalyzator (KAT). The product quality is only determined later, in another process. The interval between the product samples is between half and four hours. The melt index (MI) and the density of the polymer (ro) are monitored by off-line laboratory analysis after drying and extrusion of the polymer that causes one-hour time-delay.

The problem is to reveal the relations between the product quality and the process variables. The major aims of monitoring plant performance are the reduction of off-specification production, the identification of important process disturbances and the early warning of process malfunctions or plant faults. Furthermore, when a reliable model is available that is able to estimate the quality of the product, it can be inverted to obtain the suitable operating conditions required for achieving the target product quality.

This can be done via clustering the quality and state variables. A set of transition-free data is used that covers the whole range of specifications of the quality properties and the process variables. This data has been extracted from an SQL database. As one of the objectives is to infer the values of product quality from process data obtained at different operating regions, a set of transition-free data is used that covers the whole range of specifications of the quality properties and the process variables over all the possible operating regions. The data hold nine product grades with the following distribution

Table 2. Sample distribution of products

Code of product grade	*Number of samples*
1	2
2	76
3	5
4	1
5	65
6	94
7	103
8	11
9	52

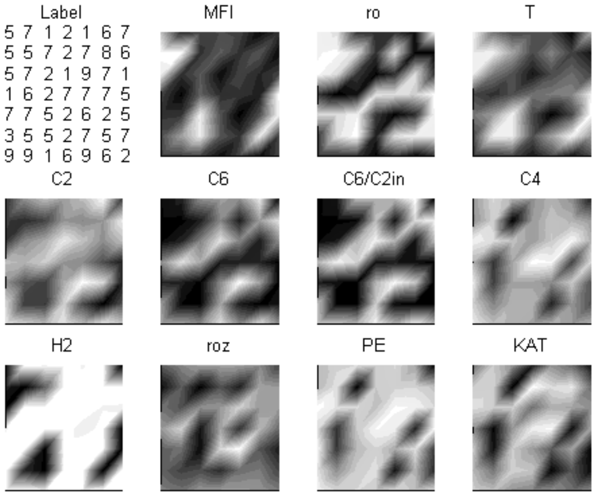

Figure 3. Code vectors obtained by the FCM algorithm

First we tried out the basic FCM algorithm. We searched 7x7=49 code vectors. The maps of the code vectors coordinates are shown on Figure 3. The values of the codebook vectors color-coded. The small values are darker the greater ones brighter and interpolated shading applied on the grid. Each cluster is labeled with the label that most of the data assigned by the greatest membership value to that partition have.

From this figure the following relations can be detected:

- The hydrogen concentration and temperature determine the melt index (MFI)
- The hydrogen , C6-C2 concentrations, C6/C2 ratio, temperature have an effect on the density of polymer
- Production intensity is influenced by the flowrate of C4 and catalyst

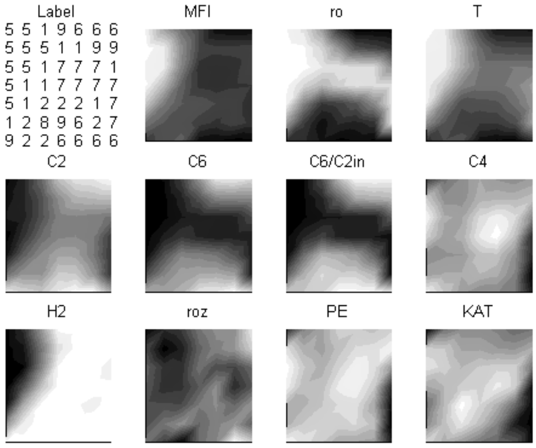

Figure 4. Map of code vectors obtained by random initialized regularized FCM

It is hard to get the information of the maps shown in Figure 3., because there are many scaterred regions with different colors, so we can’t be sure that we get all the available information by visual perception.

The regularized FCM algorithm with random initialization shows better results (see Figure 4.). Here we can see the additional information that the C2, C6 concentration influence the melt index. The relations shown in the maps of Figure 4. are easier to interpret, because there are only a few regions with different colors.

Last we initialized the code vectors on the main component plane of the data by Principal Component Analysis, and applied the regularized FCM algorithm. The resulting images (Figure 5.) are more easily interpretable as before, and similar products become much more to each other, although additional information didn’t revealed. In this figure it can be seen that although there were 9 types of products, the regularized FCM algorithm found only the Products 5, 7, 6, 9 in a well-defined region on the lattice. It is interesting to analyze why the region of product 2,7 is splitted into two regions on the map.

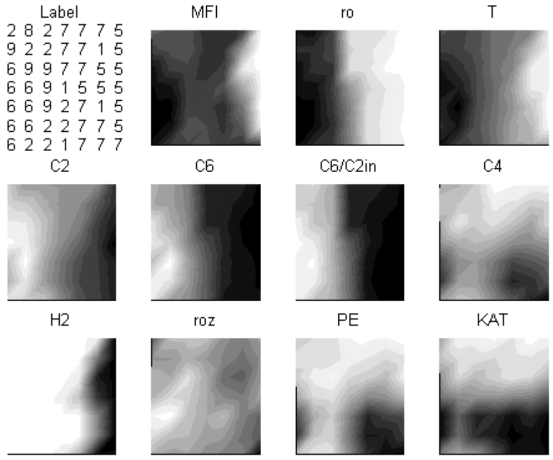

Figure 5. Map of cluster prototypes obtained by regularized FCM with initialization

For comparison the maps obtained with Kohonen's SOM can be seen on Figure 6. On this map the color codes are the negatives of ours. On the SOM map the quality variables melt index and density are the same functions of process variables as on our lattice. The border of different regions are sharper on The SOM map because we used interpolated shading and SOM presents the code vectors with cells colored proportional to their magnitude. The results show that our mapping is similar to the SOM. The FCM based algorithms found the same clusters as Kohonen's SOM and the same relations reveal.

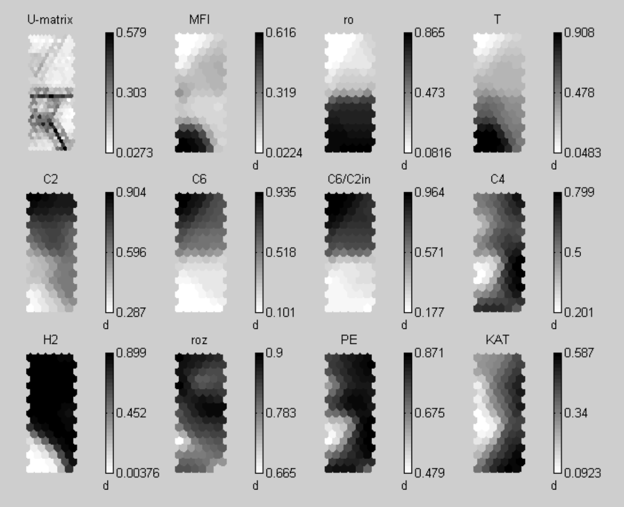

Figure 6. SOM map of the process data

4 Conclusion

With the help of clustering we were able to detect hidden relations among the data. Although, in high-dimensional problems, it is hard to interpret the results. Hence, it is extremely useful to arrange the clusters into a low dimensional grid and visualize them. This paper proposed an algorithm for this visualization task. The introduced regularization orders similar code vectors closer to each other. The proposed approach is applied in the analysis of an industrial polyethylene plant. The results show that the proposed algorithm is able to detect the relations between the process variables and the quality of the manufactured polymer.

Acknowledgements

The authors would like to acknowledge the support of the Co-operative Research Center at the University of Veszprem, (VIKKK) (KKK-II-1/A), and founding from the Hungarian Ministry of Education (FKFP-0073/2001). Janos Abonyi is grateful for the financial support of the Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences and OTKA (Hungarian National Research Foundation), No. T037600. The support of our industrial partners at TVK Ltd., especially Miklós Németh, Lóránt Bálint and Gábor Nagy is gratefully acknowledged

References

[1] T. Kohonen The Self-Organizing Map, *Proceedings of the IEEE*, 78(9) (1990), 1464-1480

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.

[3] Petri Vuorimaa, Tarko Jukarainen, Esko Kärpänoja, "A Neuro-Fuzzy System for Chemical Agent Detection", IEEE Trans. *On Fuzzy Systems*, vol 3, no. 4 Nov. 1995.

[4] A. Ahalt, A.K. Khrisnamurthy, P.Chen, D. E. Melton, "Competitive learning algorithms for vector quantization," *Neural Networks*, vol3, pp. 277-290, 1990.

[5] E. Chen-Kuo Tsao, J.C. Bezdek, N.R. Pal, Fuzzy Kohonen clustering networks, *Pattern Recognition* 27 (5), 757-764, 1994

[6] R.D Pascal-Marqui, A.D. Pascual Montano, K. Kochi, J.M.Carazo,"Smoothly distributed fuzzy c-means: a new self organizing map, *Pattern Recognition* vol.34 pp. 2395-2402 2001.

[7] Hong Yan, "Fuzzy Curve-Tracing Algorithm*", IEEE Trans. Syst., Man, Cybern.* B vol. 31, no.5. pp. 768-773 Oct. 2001.

[8] Sándor Migály, János Abonyi, Ferenc Szeifert " Fuzzy Self-Organizing Map based Regularized Fuzzy c-means Clustering " 7th-online *World Conference on Soft Computing in Industrial Applications 2002,*