Pergamon

Cement and Concrete Research 32 (2002) 1325–1330

CEMENT AND CONCRETE RESEARCH

Trace elements in clinker
II. Qualitative identification by fuzzy clustering

Ferenc D. Tamás[a,*], János Abonyi[b], János Borszéki[c], Pál Halmos[c]

[a]Department of Silicate and Materials Engineering, University of Veszprém, P.O.B. 158, H-8201 Veszprém, Hungary
[b]Department of Process Engineering, University of Veszprém, P.O.B. 158, H-8201 Veszprém, Hungary
[c]Department of Analytical Chemistry, University of Veszprém, P.O.B. 158, H-8201 Veszprém, Hungary

Received 11 July 2001; accepted 11 March 2002

Abstract

The trace element content of clinkers (and possibly of cements) can be used for the qualitative identification (i.e., manufacturing factory). This paper proposes a fuzzy classifier for the discrimination of clinkers produced in different factories based on their Mg, Sr, Ba, Mn, Ti, Zr, Zn and V content. The fuzzy classifier is identified by unsupervised fuzzy clustering. The most relevant trace elements were selected based on the obtained clusters by the modified version of the Fisher interclass separability method. The classification of a country from the European Community and South African clinkers is used as an illustrative example. The results show that the proposed method is useful to identify compact classifiers that are able to determine the origin of the clinker; the obtained classifier is easy to use and interpret for engineers and researchers, even when they are not familiar with the concept of fuzzy logic. © 2002 Published by Elsevier Science Ltd.

Keywords: Characterisation; Clinker; Trace elements; Clustering; Fuzzy classifier

1. Introduction

In our previous paper [1], the dactylogrammatic value of trace elements was described, jointly with detailed data on sample preparation and analysis; averages and standard deviations of eight trace elements (Mg, Sr, Ba, Mn, Ti, Zr, Zn and V) were tabulated. Based on >200 samples, a "standard" trace element content was calculated and, in order to facilitate the visualisation of the trace element content, a graphical method ("star plotting") was presented, where every clinker is compared to the proposed standard.

This paper pursues the development of different tools for the qualitative identification of clinkers with a study about how mathematical tools can be used for this purpose. Advanced statistical methods, called "pattern recognition" or "fingerprinting," can help qualitative identification [2]. Hierarchical clustering techniques have been applied for the clustering of Hungarian [3] and Austrian [4] clinkers, where the analytical data are transformed by principal component analysis and dendograms are constructed for cluster formation.

In this paper, a different approach is proposed that is able to identify an easily interpretable rule-based expert system. Rule-based expert systems are often applied to classification problems in fault detection, biology, medicine, etc. Fuzzy logic improves classification and decision support systems by allowing the use of overlapping class definitions and improves the interpretability of the results by providing more insight into the classifier structure and decision-making process.

Fuzzy classification techniques represent a useful and powerful tool in analytical chemistry [5]. Although several applications of fuzzy clustering and classification have been reported [6], many potential chemometric relevant applications are still waiting to be exploited.

The automatic determination of fuzzy classification rules from data has been approached by different techniques: neuro-fuzzy methods [7], GA optimization and fuzzy clustering [8].

In this paper, we assume that during the training of the classifier the factories of the samples are not known.

* Corresponding author. Tel.: +36-88-422-022x4354; fax: +36-88-423-091.

E-mail address: tamasf@almos.vein.hu (F.D. Tamás).

PII: S0008-8846(02)00797-4

Hence, clustering is applied; it is among unsupervised learning methods since it does not use a priori class identifiers. The aim of cluster analysis is the classification of objects according to similarities among them and organization of data into groups. 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. Rather, they are assigned to membership degrees between 0 and 1, indicating their partial memberships.

Among several clustering algorithms, in this paper the modified version of Gath–Geva clustering is used [9]. A method to identify a rule-based classifier from the obtained clusters is proposed, where the identified classifier is identical to the Bayes classification rule.

Using too many features results in difficulties in the prediction and interpretability capabilities of the model due to redundancy, noninformative features and noise. Hence, feature selection is necessary. For this purpose a feature selection method is proposed that is similar to the Fischer interclass separability method, which is based on statistical properties of the labelled data [10].

The rest of this paper is organised as follows. Section 2 deals with the description of the clustering algorithm, and shows the structure, the identification and the reduction of the classifier. In Section 3, a factual example of the clustering and classification of South African and European clinkers is given. This example illustrates that the proposed method is useful to identify compact and interpretable classifiers that are able to determine the origin of the clinker.

2. Identification of the fuzzy classifier

2.1. Fuzzy clustering

Cluster analysis organises data into groups according to similarities among them. In metric spaces, similarity is defined by means of distance based upon the length from a data vector to some prototypical object of the cluster. The prototypes are usually not known beforehand, and are sought by the clustering algorithm simultaneously with the partitioning of the data.

In our research work, the clustering of the trace element content is considered. Hence, the data are the observations, the trace element contents of the clinkers. Each observation consists of n measured variables, grouped into an n-dimensional column vector $z_k=[z_{1k},\ldots,z_{nk}]^T$, $z_k \in \geq \mathrm{R}^n$. A set of N observations is denoted by $\mathbf{Z}=[z_1,\ldots,z_N]$ and represented as an $n \times N$ matrix. In pattern recognition terminology, the columns of $\mathbf{Z}$ are called patterns or objects, the rows are called the features or attributes, and $\mathbf{Z}$ is called the pattern matrix.

The objective of clustering is to divide the data set $\mathbf{Z}$ into $\boldsymbol{c}$ clusters. A $c \times N$ matrix $\mathbf{U}=[\mu_{ik}]$ represents a fuzzy partition if its elements satisfy the following conditions:

$$\mu_{ik} \in [0,1], \quad \sum_{i=1}^{c} \mu_{ik} = 1, \quad 0 < \sum_{i=1}^{c} \mu_{ik} < N,$$
$$1 \leq i \leq c, \quad 1 \leq k \leq N \tag{1}$$

where c is the number of the fuzzy clusters, μ_{ik} denotes the degree of the membership, how the $\boldsymbol{z}_k=[z_{1k},\ldots,z_{nk}]^T$th observation belongs to the $1 \leq i \leq c$th cluster.

The objective of the fuzzy clustering is to minimize the sum of the weighted squared distances between the data points and the cluster prototypes that are in this case the centers of the clusters

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

where $\mathbf{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_c\}$ is the set of vectors of the cluster prototypes (centers), $m \in (1, \infty)$ is a weighting exponent that determines the fuzziness of the resulting clusters and is chosen as $m=2$.

The minimization of the functional Eq. (2) subjected to the constraints of Eq. (1) represents a nonlinear optimization problem that can be solved by using a variety of available methods. The most popular method, however, is alternating optimization (AO), which is given in Table 1.

2.2. Structure of the classifier

We apply fuzzy classification rules that each describe one of the c classes in the data set. The rule antecedent is a fuzzy description in the n-dimensional feature space and the rule consequent is a class label from the set $\{1,2,\ldots,c\}$:

$$R_i \qquad \text{If } z_{1,k} \text{ is } A_{i,1}(z_{1,k}) \text{ and } \ldots\ z_{n,k} \text{ is } A_{i,n}(z_{n,k})$$
$$\text{Then } l_i \qquad [w_i] \tag{3}$$

Here, the label of the ith cluster, l_i, is the output of the ith rule, and $A_{i,j}$ is a linguistic term (for instance "Big") defined by a fuzzy set. In this paper, Gaussian membership functions are used to represent the fuzzy sets (Eq. (4)),

$$A_{i,j}(z_{j,k}) = \exp\left(-\frac{1}{2}\frac{(z_{j,k} - v_{j,k})^2}{\sigma_{j,i}^2}\right) \tag{4}$$

where $\sigma_{i,j}^2$ represents the variance of the Gaussian function.

In Eq. (3), $w_i \in [0,1]$ is the weight of the rule that represents the desired impact of the rule. The value of w is often chosen by the designer of the fuzzy system based on his or her belief in the goodness and accuracy of the ith rule. When such knowledge is not available w_i is set as $w_i=1$, $i=1,\ldots,c$.

The "and" connective is modelled by the product operator, allowing for interaction between the propositions in the

Table 1
Algorithm of the clustering procedure

Initialization:
Given the data set $\mathbf{Z}$, choose the number of clusters c, the weighting exponent $m=2$, the termination tolerance $\varepsilon>0$ and initialize the partition matrix randomly.
Repeat for $l=1,2,\ldots$
Step 1: Compute the cluster centers:

$$v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m z_k}{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m}, \; 1 \le i \le c$$

Step 2: Compute the fuzzy $\mathbf{F}_i$ covariance matrix and the P_i prior probability

$$\mathbf{F}_i = \frac{\sum_{k=1}^{N} \mu_{ik}(z_k - v_i)(z_k - v_i)^T}{\sum_{k=1}^{N} \mu_{ik}}$$

$$P_i = \frac{1}{N}\sum_{k=1}^{N} \mu_{ik}$$

Step 3: Compute the distances

$$D_{ik}^2 = 1 \bigg/ \left(\frac{P_i}{(2\pi)^{n/2}\sqrt{\det \mathbf{F}_i}} \exp\left(-\frac{1}{2}(z_k - v_i)^T \mathbf{F}_i^{-1} (z_k - v_i) \right) \right)$$

Step 4: Update the partition matrix:

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

otherwise $\mu_{ik}^{(l)}=0$
until $\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\| < \varepsilon$

antecedent. The degree of activation of the ith rule is calculated as:

$$\beta_i(\mathbf{z}_k) = w_i \prod_{j=1}^{n} A_{i,j}(z_{j,k}) \tag{5}$$

The output of the classifier is determined by the rule that has the highest degree of activation:

$$y_k = l_{i^*} i^* = \operatorname{argmax}_i \beta_i(\mathbf{z}_k) \tag{6}$$

2.3. Constructing the classifier from clusters

The aim of this subsection is to show how fuzzy clustering can be used for the identification of the classifier presented above.

The use of Gaussian membership function allows the compact formulation of Eq. (5) (Eq. (7))

$$\beta_i(\mathbf{z}_k) = w_i \exp\left(-\frac{1}{2}(\mathbf{z}_k - \mathbf{v}_j)^T \mathbf{F}_i^{-1} (\mathbf{z}_k - \mathbf{v}_j) \right) \tag{7}$$

where $\mathbf{F}_i$ matrix contains the variances of the Gaussian membership functions (Eq. (8)):

$$\mathbf{F}_i = \begin{bmatrix} \sigma_{1,i}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2,i}^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{n,i}^2 \end{bmatrix} \tag{8}$$

When the correlations among the input variables are not taken into account during clustering, only the diagonal elements of $\mathbf{F}_i$ are calculated in Step 2 of the algorithm shown in Table 1. In this case, the clusters obtained by fuzzy clustering can be used directly to determine the $\sigma_{j,i}^2$ parameters of the membership functions and the rule weights (Eq. (9)):

$$w_i = \frac{P_i}{(2\pi)^{n/2}\sqrt{\det(\mathbf{F}_i)}} \tag{9}$$

The fuzzy classifier defined by the (Eq. (6)) decision rule is identical to the classic Bayes classifier when the parameters of the fuzzy system are obtained by the method presented above.

2.4. Feature selection based on interclass separability

The interpretability of the classifier depends on the number of the utilized features. For the selection of the most relevant features, we modify the Fischer interclass separability method, which is based on statistical properties of labeled data [9]. The interclass separability criterion is based on the *between-class* and *within-class* covariance matrices that sum up to the *total* covariance of the training data $\mathbf{F}_T = \boldsymbol{F}_B + \boldsymbol{F}_W$, where (Eq. (10))

$$\mathbf{F}_w = \sum_{i=1}^{c} P_i \mathbf{F}_i \;\; \mathbf{F}_B = \sum_{i=1}^{c} P_i (\mathbf{v}_0 - \mathbf{v}_i)(\mathbf{v}_0 - \mathbf{v}_i)^T$$

$$\text{where } \mathbf{v}_0 = \sum_{i=1}^{c} P_i \mathbf{v}_i \tag{10}$$

The feature interclass separability selection criterion is a trade-off between $\mathbf{F}_B$ and $\mathbf{F}_W$ (Eq. (11)):

$$J = \frac{\det \mathbf{F}_B}{\det \mathbf{F}_w} \tag{11}$$

The importance of a feature is measured by leaving out the feature and calculating J for the reduced covariance

Table 2
Trace element content of European clinkers (mg/kg)

Code	Ba	Mn	Sr	Ti	Zr	Mg	V	Zn	Factory
EU 1	154	248	570	1,234	39	7,655	115	78	1
EU 7	144	212	546	1,245	41	5,760	103	85	1
EU 14	130	213	520	1,204	39	5,873	118	91	1
EU 22	129	229	569	1,247	32	6,490	98	71	1
EU 2	165	154	248	1,414	48	5,804	38	301	2
EU 9	134	132	210	1,391	47	5,272	35	267	2
EU 15	136	134	227	1,313	49	5,519	34	428	2
EU 24	124	130	212	1,378	48	5,421	33	257	2
EU 3	125	205	397	1,550	74	13,873	69	76	3
EU 8	109	173	404	1,650	74	13,070	100	74	3
EU 13	96	168	396	1,502	66	14,700	83	67	3
EU 23	114	177	378	1,604	58	12,987	71	72	3
EU 4	82	218	249	1,342	46	10,737	127	168	4
EU 12	84	206	230	1,446	58	11,752	135	298	4
EU 16	79	227	261	1,444	53	18,353	165	216	4
EU 17	77	225	271	1,530	60	18,042	155	157	4
EU 5	159	285	146	846	35	10,314	114	192	5
EU 11	117	211	497	1,282	45	8,997	30	227	5
EU 18	110	186	410	1,361	49	9,264	31	194	5
EU 19	72	164	414	1,178	44	8,196	124	78	5
EU 6	239	217	488	1,232	45	9,197	106	134	6
EU 10	126	305	141	806	34	9,954	52	153	6
EU 20	123	231	178	847	36	9,276	31	129	6
EU 21	146	211	209	956	46	10,406	31	90	6

matrices. The feature selection is made iteratively by leaving out the less needed feature.

3. Experimental

3.1. Materials and analytical experiments

For the qualitative "fingerprinting" of clinkers, obviously a set of well-defined clinker samples are necessary: composite average samples of a longer period of kiln operation. A nationwide sample collection has a limited value only; to obtain a higher area, a Technical Committee Qualitative Identification of Clinkers and Cements (QIC) has been established in 1996 under the auspices of Réunion Internationale des Laboratoires d'Essais et de Recherches sur les Matériaux et les Constructions (RILEM) (TC 180/QIC). The details of the collection process and analytical experiments are described in our previous paper [1].

3.2. Application of fuzzy clustering and classification

In this paper, clinkers from South Africa and from a European country[1] are analysed. Twenty and 24 samples come from five and six South African and European factories, respectively. These samples have been analysed to determine their Ba, Mn, Sr, Mg; Ti, Zr, Zn and V content. The first six elements come from the main raw materials and are of dactylogrammatic value, while the last two elements mainly come from the fuel (used tires, heavy fuel oil, etc.) and cannot be used for identification. The trace element content of the European and South African clinkers are shown in Tables 2 and 3, respectively.

As fuzzy clustering is an unsupervised learning method, during the identification of the fuzzy classifier, the class labels (factories, last column of Tables 2 and 3) were unknown.

The clustering algorithm is sensitive to variations in the numerical ranges of different features. Hence, the performance of the obtained classifier can be negatively influenced by the different magnitude of the trace element contents. Therefore, the clustering was performed based on normalised data, where all transformed features have zero mean and unit variance (Eq. (12)),

$$\tilde{z}_{j,k} = \frac{z_{j,k} - \bar{z}_j}{\sigma_j} \tag{12}$$

where $\bar{z}_j$ represents the mean and σ_j the variance of the jth feature.

The proposed fuzzy clustering and classification method has been implemented in MATLAB. The program can be downloaded from the homepage of author J.A: www.fmt.vein.hu/softcomp.

Two fuzzy classifiers were identified, one for the South African and one for the European clinkers.

The model accuracy is measured in terms of the number of misclassifications. Table 4 shows the performance of the fuzzy classifiers.

Later information, revealed by the sample supplier (a member of RILEM TC 180-QIC) confirmed the accuracy of the method, with two and two exceptions, respectively, in case of the European and South African samples, meaning

Table 3
Trace element content of South African clinkers (mg/kg)

Code	Ba	Mn	Sr	Ti	Zr	Mg	V	Zn	Factory
SA 1	146	428	1,024	853	18	2,693	20	9	1
SA 10	155	451	1,192	943	30	3,165	20	9	1
SA 16	168	444	1,168	893	0	3,298	22	19	1
SA 22	235	386	2,110	1,061	54	5,241	23	106	1
SA 2	569	3,003	49	1,178	32	15,265	47	39	2
SA 9	604	2,933	19	1,242	63	15,838	44	13	2
SA 3	485	7020	213	1,052	41	24,292	29	12	3
SA 8	558	6,566	179	1,126	73	23,225	26	11	3
SA 17	407	6,638	164	1,136	33	24,125	26	18	3
SA 20	449	4,519	168	957	68	20,973	29	29	3
SA 4	207	584	2,090	25	30	5,356	27	26	4
SA 5	210	610	2,126	1,176	9	5,358	26	25	4
SA 12	195	509	2,296	1,224	18	5,659	24	26	4
SA 13	193	490	2,298	1,252	45	5,420	24	28	4
SA 15	191	497	2,274	1,208	32	5,523	24	30	4
SA 21	176	379	2,142	1,102	56	5,039	23	29	4
SA 23	174	434	1,058	880	50	3,126	19	21	4
SA 6	122	264	2,934	804	15	5,680	17	40	5
SA 11	136	210	3,107	903	47	6,723	19	14	5
SA 19	165	491	3,484	805	31	6,314	20	40	5

[1] Country name withheld by request.

Table 4
Performance of fuzzy classifiers

	European clinkers			South African clinkers		
Factory	No. data	No. miss.	Performance, %	No. data	No. miss	Performance, %
1	4	0	100	4	1	75
2	4	0	100	2	0	100
3	4	0	100	4	0	100
4	4	0	100	7	1	85.71
5	4	1	75	3	0	100
6	4	1	75			
Total	24	2	91.67	20	2	90

that the performance of the classifiers (Table 4) is in case of the investigated 44 clinker samples $\geq 90\%$.

The exceptions (real factory/factory determined by fuzzy clustering), are in case of European clinkers EU 5 and EU 6 and in case of South African clinkers SA 22 and SA 23.

It is interesting to note that when the Zn and V content of the clinkers were also used to identify the classifier, the classification performance become worse. For European clinkers, the clustering algorithm was not able to detect the group of clinkers produced in Factory 5. For South African clinkers the number of misclassifications has increased to three, the exceptions become SA 21, SA 22 and SA 23. This result shows, as the Zn and V do not have dactylogrammatic value (they mainly come from the fuel) and they cannot be used for the identification of clinkers.

It can happen that the obtained fuzzy model is unnecessarily complex as not all the features are needed to identify the clinker. Hence, the feature selection algorithm presented in Section 2.4. has been used to iteratively leave out the unnecessarily features. The reduced fuzzy classifier utilises three features (Sr, Mg and Mn element contents) and gives the same performance as the original system. For the correct classification of European clinkers, the Sr, Mg and Ti concentrations were needed.

The obtained classifier is easy to use and interpret for engineers and researchers, even when they are not familiar with the concept of fuzzy logic. For instance, the first rule of the classifier of European clinkers is

If Sr is *around* 250 and Ti is
around 1370 and Mg is *around* 5500 Then Factory 1

where the meaning of the linguistic terms "around" can be seen in Fig. 1.

4. Conclusions

The trace element content of clinkers (and possibly of cements) can be used for the qualitative identification (i.e., determination of the manufacturing plant). For this purpose, several samples from South Africa and Europe have been analysed to determine their Ba, Mn, Sr, Mg, Ti, Zr, Zn and V content. The first six elements come from the main raw materials and are of dactylogrammatic value, while the last two elements mainly come from the fuel (used tires, heavy fuel oil, etc.) and cannot be used for identification.

For the qualitative identification of clinkers a rule-based classifier was designed trained by unsupervised fuzzy clustering. The most relevant trace elements were selected based

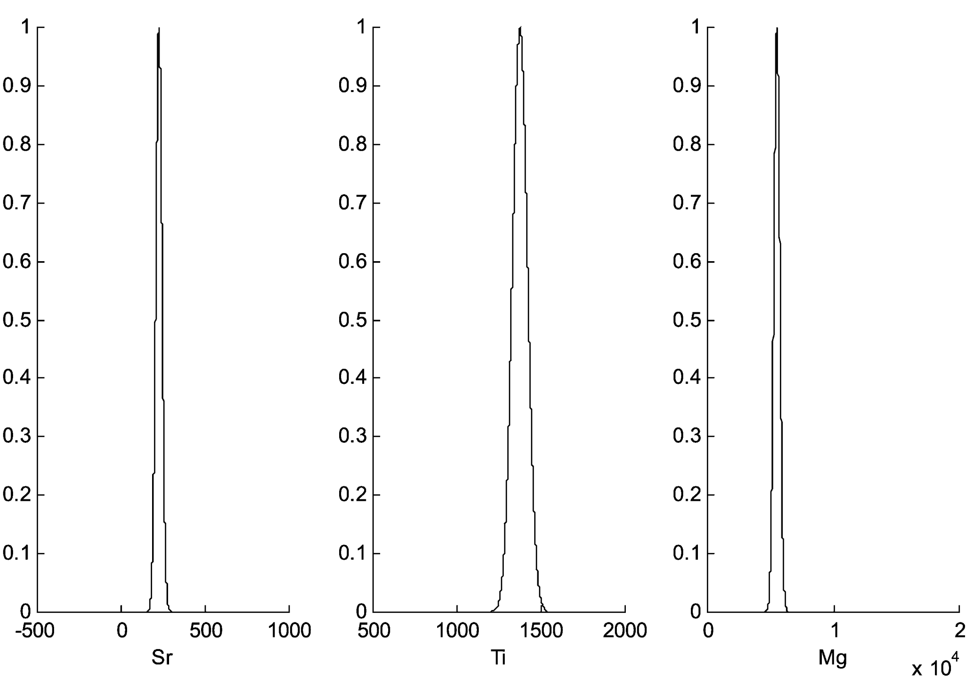

Fig. 1. Membership functions of the first rule of the fuzzy classifier designed to categorize European clinkers. The trace element content is given in milligrams per kilogram.

on the obtained clusters by the modified version of the Fisher interclass separability method. It has turned out that Ba, Mn, Mg are the most relevant trace elements for fingerprinting.

The results show that the proposed method is useful to identify compact classifiers that are able to determine the origin of the clinker. The obtained classifier is easy to use and interpret for engineers and researchers, even when they are not familiar with the concept of fuzzy logic.

A detailed description helps the implementation of the clustering algorithm; still easier, a program has been constructed and can be downloaded (www.fmt.vein.hu/softcomp).

Acknowledgments

The financial support of OTKA (Hungarian National Research Foundation), No. T026307 is gratefully acknowledged. Thanks are due to the laboratories of Hungarian cement factories as well as to members of the Technical Committee ‘‘180-QIC’’ (Qualitative Identification of Clinkers and Cements) of Réunion Internationale des Laboratoires d’Essais et de Recherches sur les Matériaux et les Constructions (RILEM) for collecting composite average clinker samples. János Abonyi is grateful for the Janos Bolyai Research Fellowship of the Hungarian Academy of Science.

References

[1] F.D. Tamás, J. Abonyi, Trace elements in clinkers: I. A graphical representation, Cem. Concr. Res. 32 (2002) 1319–1323.

[2] J.C. Miller, J.N. Miller, Pattern recognition, Statistics for Analytical Chemistry, New York, Ellis Horwood, 1984 (Chapter 7.13).

[3] F.D. Tamas, Pattern recognition methods for the qualitative identification of Hungarian clinkers, World Cem. Res. Dev. 27 (1996) 75–79.

[4] F.D. Tamas, A. Tagnit-Hamou, J. Tritthart, Trace elements in clinker and their use as ‘‘fingerprints’’ to facilitate their qualitative identification, in: M. Cohen, S. Mimdess, J. Skalny (Eds.), Materials Science of Concrete: The Sidney Diamond Symposium, The American Ceramic Society, Westerville, OH, 1998, pp. 57–69.

[5] D.H. Rouvray (Ed.), Fuzzy Logic in Chemistry, Academic Press, San Diego, CA, 1997.

[6] G. Barkó, J. Abonyi, J. Hlavay, Application of fuzzy clustering and piezoelectric chemical sensor array for investigation on organic compounds, Anal. Chim. Acta 398 (2–3) (1999) 219–222.

[7] J.S.R. Jang, C.T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing; a Computational Approach to Learning and Machine Intelligence, Prentice-Hall, Upper Saddle River, NJ, 1997.

[8] M. Setnes, J.A. Roubos, GA-fuzzy modeling and classification: complexity and performance, IEEE Trans. Fuzzy Syst. 8 (5) (2000) 509–522.

[9] F. Hoppner, R. Kruse, T. Runkler, Fuzzy Cluster Analysis—Methods for Classification, Data Analysis and Image Recognition, Wiley, New York, 1999.

[10] J.A. Roubos, M. Setnes, J. Abonyi, Learning fuzzy classification rules from data, in: R. John, R. Birkenhead (Eds.), Developments in Soft Computing, Springer-Verlag, Berlin, 2001, pp. 108–115.