# Analysis of Trace Elements in Clinker based on Supervised Clustering and Fuzzy Association Rule Mining

F. D. Tamás, F. P. Pach, J. Abonyi
*University of Pannonia, Veszprém, Hungary*

Abstract

Trace element content of clinkers is of high scientific interest, and can be used to solve practical problems too. In this paper qualitative identification to determine the origin of clinkers is described. The classification of clinkers produced in different factories can be based on their trace element content. A new association rule based method has been proposed to identify an interpretable classification system. The performance of the obtained fuzzy rule base classifier was measured by ten-fold cross validation. The results show that the proposed method is useful to identify easy-to-use expert systems that are able to determine the origin of the clinker based on its trace element content.

Keywords: clinker, trace element, classification, clustering, association rule mining

## 1. Introduction

Trace element content of clinkers can be used to determine the origin of the clinker (i.e. the manufacturing factory). The first paper of similar topics was published in 1993 [1]. This first attempt suggests that pattern recognition or fingerprinting can help qualitative identification [2]. However, the qualitative identification obviously requires a database, to compare the trace element content of unknown clinkers/cements with characteristic known samples. Data, describing trace element content of clinkers and cements have been published too [3, 4, 5]. In these papers it was shown, that not all trace elements can be used for fingerprinting; selection must follow certain principles. The most important item of selection: trace elements of “dactylogrammatic value” should come from the main raw material (limestone, marl, clay) and not from the fuel, from furnace lining or from grinding media wear, and some other principles should be observed as well. More recently 6 elements were used to characterize clinkers: besides those used in [1] the Mg, Ti and Zr contents were measured too [4, 5]. Zn and V have no dactylogrammatic value (they come from the fuel, if waste tyres or special sorts of heavy fuel oil are used, resp.), but their quantity can be interesting in cement performance.
In our previous paper [6] 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 more than 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. Among the wide range of advanced statistical and pattern recognition methods, hierarchical clustering technique have been applied for the clustering of clinkers, where the analytical data were transformed by principal component analysis and dendograms were constructed for cluster formation [3, 4, 5].

Rule-based expert systems are often applied to classification problems in fault detection, biology, medicine, etc. In [7] a new approach has been proposed to identify an interpretable fuzzy rule-based expert system. 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 [8]. A chemometric example for the effective use of fuzzy clustering can be found in [9].

In our previous work [10] a fuzzy clustering and decision tree based method has been introduced to identify compact and accurate classifiers that are able to efficient determine the origin of the clinker.

This paper introduces a new method to generate interpretable rule based expert system to clinker classification problem. In our method the classifiers can be formulated by fuzzy association rules discovered on the collected data. In the last decade many associative classification algorithms have been developed [11, 12, 13]. Classification model generated by associative methods contrast with black-box techniques are understandable to humans. But most of the methods have a disadvantage, namely interpretability of the classifier (rule base) is neglected besides the accuracy, because the prediction is based on hundreds of rules. It induces problems because in many (chemical engineering) applications the understanding of data and the classification problem is critical importance. The ideal thing would be to satisfy accuracy and interpretability together to a high degree, but since they are contradictory issues, it is generally not possible. Our main goal is that select a compact set of classification rules which has good interpretability and then determine an efficient classification model to get high classification accuracy. To achieve our goal a new fuzzy associative classification algorithm is developed. The rest of this paper is structured the following. The basic definitions of the classification and the association rule mining are introduced in Section 2. Main steps of the proposed method are presented in Section 3. The new method is used to determine the origin (i.e. manufacturing factory) of clinkers in a European country.

## 2. Classification and association rule mining

*Classification* is a data mining procedure in which individual items are placed into groups (classes) based on quantitative information on one or more characteristics inherent in the items (referred to as features or variables, etc). The introduction to data mining and this topic is detailed in [15]. A typical data set of a classification problem contains numerical (continuous) and/or categorical (discrete) values (attributes). The categorical attribute that denotes the class of the item is called *class label*. The remaining attributes are called *predictor attributes*. The classification task is to find a model which can be used to determine the class label of previously unknown item based on some predictor attributes of the item.
The *association rule mining* problem originates in market basket analysis which aims at understanding the behavior of retail customers, or in other words, finding associations among the items purchased together [14]. A famous example in a supermarket database is “diapers => beer”, i.e. young fathers being sent off to the store to buy diapers, reward themselves for their trouble. A general form of *association rule* is the X => Y implication where the two parts of rule are the *antecedent* (X) and *consequent* (Y). Both part consist items (item sets). Rules with only a class label in consequent part need to mine (X => C) to classification. Therefore an associative classification task is to build a classification model (classifier) based on a set of mined *classification association rules (CAR)*.
In the conventional (crisp) logic if an attribute is divided to several sets, each data point belong to only one set (e.g. on attribute Age, a people can be young or old). In fuzzy logic all data point can belong to each set with several membership values (a people can be young and old in the same time, but with several memberships). The fuzzy sets of an attribute are called membership function (fuzzy interval). The Fig. 1 shows an example for the memberships.
In the following basic definitions of the fuzzy association rule mining is presented.

Definition: Let $A_{i,j}$ is a fuzzy interval on attribute $z_i$. Then $\langle z_i : A_{i,j} \rangle$ is called attribute-fuzzy interval pair, or simply *fuzzy item*.
An example could be $\langle \text{Age} : \text{young} \rangle$. The set of fuzzy items is called *fuzzy item set*: $\langle Z : A \rangle = [\langle z_{i1} : A_{i1,j} \rangle \cup \langle z_{i2} : A_{i2,j} \rangle \cup \ldots \cup \langle z_{iq} : A_{iq,j} \rangle]$,

where $q < n + 1$ (n denotes the number of attributes). For example in case of attributes $Z = \{Age, Height\}$ an item set can be the: $\langle \text{Age} : \text{young} \rangle \cup \langle \text{Height} : \text{small} \rangle$.

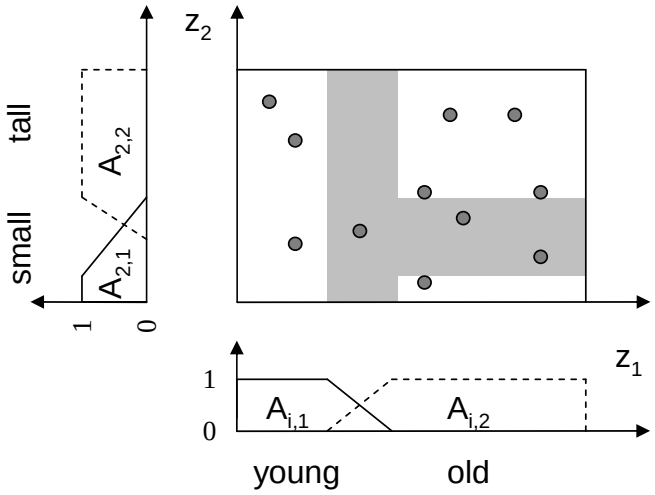


Figure 1 Example for fuzzy membership function

Definition: The *fuzzy support* of the item set reflects how the record of the identification data set support the item set. It is calculated by the following equation:

$$FS(\langle Z : A \rangle) = \frac{\sum_{k=1}^{N} \prod_{\langle z_i : A_{i,j} \rangle \in \langle Z:A \rangle} t_k(z_i)}{N}$$

where $t_k(z_i)$ denotes the *k*th fuzzy data on the *i*th attribute. The fuzzy data of a data points are the membership values for the membership functions defined on the attribute. The N denotes the number of records in the data set.

The following example illustrates the calculation of the fuzzy support value of a fuzzy item set. The example data set is shown in Table 1. Let $\langle Z : A \rangle = \langle \text{Balance} : \text{medium} \rangle \cup \langle \text{Income} : \text{height} \rangle$ is an item set, the fuzzy support of $\langle Z : A \rangle$ is calculated by:

$$FS(\langle Z : A \rangle) = \frac{0.5 \cdot 0.4 + 0.8 \cdot 0.4 + 0.7 \cdot 0.7}{3} = 0.337$$

Definition: A fuzzy item set is called *frequent* fuzzy item set if its fuzzy support value is higher than or equal to (user defined) minimal fuzzy support threshold.

| $\langle$Age : old$\rangle$ | $\langle$Balance : medium$\rangle$ | $\langle$Income : height$\rangle$ |
|---|---|---|
| 0.4 | 0.5 | 0.4 |
| 0.3 | 0.8 | 0.4 |
| 0.2 | 0.7 | 0.7 |

Table 1 Example data set containing membership values

Definition: The $\langle X : A\rangle \Rightarrow \langle Y : B\rangle$ implication is called *fuzzy association rule* where the item sets are frequent in both parts of the implication.

Definition: The *fuzzy confidence* of the association rule is calculated by:

$$FC(\langle X : A\rangle \Rightarrow \langle Y : B\rangle) = \frac{FS(\langle X : A\rangle \Rightarrow \langle Y : B\rangle)}{FS(\langle X : A\rangle)}$$

which can be understood as the conditional probability of the consequent part $\langle Y : B\rangle$, namely $P(\langle Y : B\rangle \mid \langle X : A\rangle)$.

Definition: The *fuzzy correlation measure* of the association rule describes the relationship between antecedent and consequent parts. It is calculated by:

$$FCORR(\langle X : A\rangle \Rightarrow \langle Y : B\rangle) =$$

$$\frac{FS(\langle X : A\rangle \cup \langle Y : B\rangle) - FS(\langle X : A\rangle) \cdot FS(\langle Y : B\rangle)}{\sqrt{FS(\langle X : A\rangle) \cdot (1 - FS(\langle X : A\rangle)) \cdot FS(\langle Y : B\rangle) \cdot (1 - FS(\langle Y : B\rangle))}}$$

Definition: The *firing strength* of a fuzzy rule is determined by the mechanism which is used to implement the And operation in the antecedent part of the rules. In this paper the product of degrees of membership is proposed. Therefore the firing strength of the *j*th rule for the *k*th sample is calculated by:

$$\beta_j(x_k) = \prod_i t_k(z_i), \quad \langle z_i : A_{i,j}\rangle \in \langle Z : A\rangle$$

where $z_i$ denotes the *i*th attribute.

After the basics of the fuzzy association rule mining the main steps of our new method are showed.

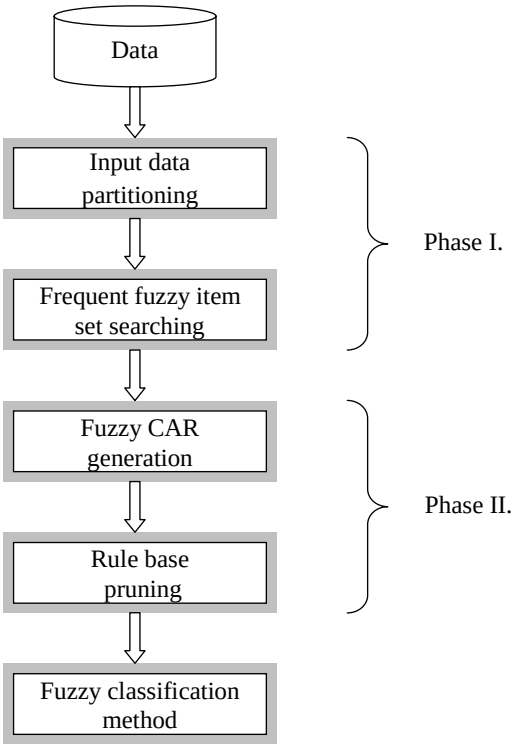


Figure 2 The proposed associative classification method

## 3. The proposed classification method

Most of associative classification algorithms find frequent item sets in one phase and discover classification rules in other phase. Our algorithm is also divided to these phases. The main steps of the method are the followings: (presented by the blocks in Fig. 2)

*Step 1*: To classification by an associative classifier, there is need to use some partition method in case of continuous data attributes to the determination of the item sets. Because the right partitioning of the continuous predictor attributes has significant effect to the classification performance a supervised fuzzy clustering algorithm is proposed instead of a crisp approach. The selected Gath-Geva algorithm is a supervised clustering method [16]. It uses the class label to determine the clusters of data. It results Gaussian membership functions to represent the fuzzy sets (items are equal to clusters) on the attributes. These membership functions are transformed into trapezoids and based on them a fuzzy data set is generated from the original crisp data set.

*Step 2*: After the input domain partition step the resulted fuzzy sets (trapezoids) are used as items. An Apriori based frequent fuzzy item set searching method is proposed which serves the input (set of item sets) for rule base generation step. Before the start of searching the threshold value, namely the minimal fuzzy support is defined (by the user). (In our

present implementation this value is set to 50 %.) Then in the first step all the possible 1-length item sets are listed (as 1-candidates). The frequent of them are used to construct the possible 2-length item sets (as 2-candidates). While frequent item sets are found this candidate construction and frequent searching is run.
*Step 3*: Classification association rules are built easy from discovered frequent item sets. All the possible rules are generated which consist only a class label as consequent. Besides the fuzzy confidence a correlation measure is used to determine importance values of rules.
*Step 4*: After classification rules are generated and structured into a fuzzy rule base, a three-step post pruning method is also applied to select the most important rules to classification. The determination of the importance of rules and the rule base pruning method are based only the (previously defined) fuzzy correlation measure.
*Step 5*: For the classification of a new sample first the product of the firing strength and the fuzzy confident value is calculated for all the rules in the rule base. The classes are scored by these product values of all the rules to consider the class labels in the rules. The class label with top score will be the predicted class.

## 4. Application study

For the qualitative “fingerprinting” of clinkers, obviously a set of well-defined clinker samples are necessary. Twenty clinker samples come from five factories of a European country. The samples have been analyzed to determine their Mg, Sr, Ba, Mn, Ti, Zr, Zn and V content.

| Trace element | Min | Max | Typical min | Typical max |
|---|---|---|---|---|
| Ba | 77 | 239 | 47 | 442 |
| Mn | 130 | 305 | 15 | 6538 |
| Sr | 141 | 570 | 19 | 2972 |
| Ti | 806 | 1650 | 175 | 1691 |
| Zr | 32 | 74 | 4 | 149 |
| Mg | 5272 | 18353 | 1751 | 24125 |
| Zn | 67 | 428 | 11 | 559 |
| V | 31 | 165 | 17 | 297 |

Table 2 Trace element content of samples (mg/kg)

The Table 2 shows the contents and typical ranges of the elements [6]. The proposed classification method is used to determine the origin (i.e. manufacturing factory) of clinkers. The key of determination is the amount of the eight trace elements, but the last two, namely the Zn and V mainly come from the fuel therefore they cannot be used for identification. Both type of data set (with and without Zn and V) are applied and the resulted classifiers are compared. The accuracy of classification models is

measured in terms of the number of misclassifications. The performances of classifiers were measured by ten-fold cross validation. This means that the data is divided into ten sub-sets of cases that have similar size and class distributions. Each sub-set is left out once, while the other nine remaining are applied for the construction of the classifier which is subsequently validated for unseen cases in the left-out sub-set.

4.1 Determination of the origin of clinker based on trace element content

In the determination test, first performances of the new fuzzy method and a frequently used classification algorithm, the C4.5 are compared [17]. The selected classification data set has five classes (the five factories). It can be easy transformed to binary classification problem, where the task is to separate samples in two classes where the first is defined by a marked class and the second is defined by the remainder four classes. Both algorithms are tested in multi class and binary classification problems too. First the identification is based on the listed eight elements the results are summarized in Table 3. The first column shows the type of classification (e.g. in first five rows, indices of binary classifications for the five factories are placed). The second and third columns show the performances of the C4.5 and the new method for the classification problems, respectively. The performances are represented by the accuracy and the complexity (number of rules and conditions are placed in brackets) of the classifiers. The results show, that our new method have higher accuracy (96 %) in mean for the binary problems, but the C4.5 generates more compact classifiers (Rule bases are contains only 2 rules and 2 conditions.). In the original problem with five classes, the proposed method classifies lots better (see the last row). The reason of this significant difference can be originated from the classification mechanism of C4.5. It is a greedy algorithm, consequently during the decision tree induction step the selection of attributes is based on the principle of the information gain. It selects the most informative attribute in the crisp cuts. It can be result that this type of method does not always select all, and right attributes. Therefore the identification based on data set without the two previously mentioned trace elements is need tested.

| **Factory** | **C4.5** | **New method** |
|---|---|---|
| 1 | 95 (2, 2) | 95 (4.3, 15.1) |
| 2 | 95 (2, 2) | 100 (4.4, 15.6) |
| 3 | 75 (2, 2) | 95 (5.1, 16.2) |
| 4 | 100 (2, 2) | 95 (4.8, 14.0) |
| 5 | 90 (2, 2) | 95 (3.8, 14.1) |
| Mean performance | 91 (2, 2) | 96 (4.48, 15) |
| Multi class (1-5) | 65 (5, 10.9) | 90 (6.3, 10.3) |

Table 3 Performances of the C4.5 and our fuzzy classifiers

| Factory | C4.5 | New method |
|---|---|---|
| 1 | 95 (2, 2) | 95 (3.2, 6.8) |
| 2 | 90 (2, 2) | 100 (4, 10.5) |
| 3 | 80 (2, 2) | 95 (3.7, 7.5) |
| 4 | 95 (2, 2) | 95 (2.6, 5.8) |
| 5 | 90 (2, 2) | 95 (2.1, 3.5) |
| Mean performance | 90 (2,2) | 96 (3.12, 6.82) |
| Multi class (1-5) | 85 (5, 10.7) | 85 (6.6, 12.1) |

Table 4 Performances of the C4.5 and our fuzzy classifiers (identification is based on data set without Zn and V)

The results are summarized in Table 4 where the structure is the same as was in Table 3. While the mean accuracy of the proposed method is the same as previously, the classifiers are more compact (rules: 4.48 -> 3.12, conditions: 15 -> 6.82). The increasing of the accuracy of C4.5 classifier is notable in the multi class problem (65 % -> 85 %). To the detailed analysis of classifiers, the structure of the generated rule bases can be compared.

4.2 Structure of classification rule bases, determination of the most important trace elements

Structures of generated rule bases can be easy compared in case of associative classifiers, because the rule bases are good interpretable for humans. The Fig. 3 shows an example for a fuzzy classification rule base generated by supervised clustering based approach. Each row of the figure contains a classification rule. The first six columns contains the antecedent parts of the rules, the last column represents the consequents, the class labels (index of factories). For example, the first rule of the classifier of clinkers is:

"**If** Mn content is *around* 130-181 **and** Zr content is *around* 38-54
**Then** Factory 2"

The fuzzy correlation value of this rule is near to the one (0.9), therefore this rule is highly important in determination of the clinkers produced in the second factory. During the rule searching method these important rules are selected. The final rule base (e.g. the represented in Fig. 3) is built from them. While performances of classifiers were measured by ten-fold cross validation it means that ten fuzzy rule bases are generated during the validation. Because the training data is not the same in each validation cycle the rule bases can be different. To get information about trace elements which are the most determine in prediction of classes, the aggregation of the rule bases is necessary.

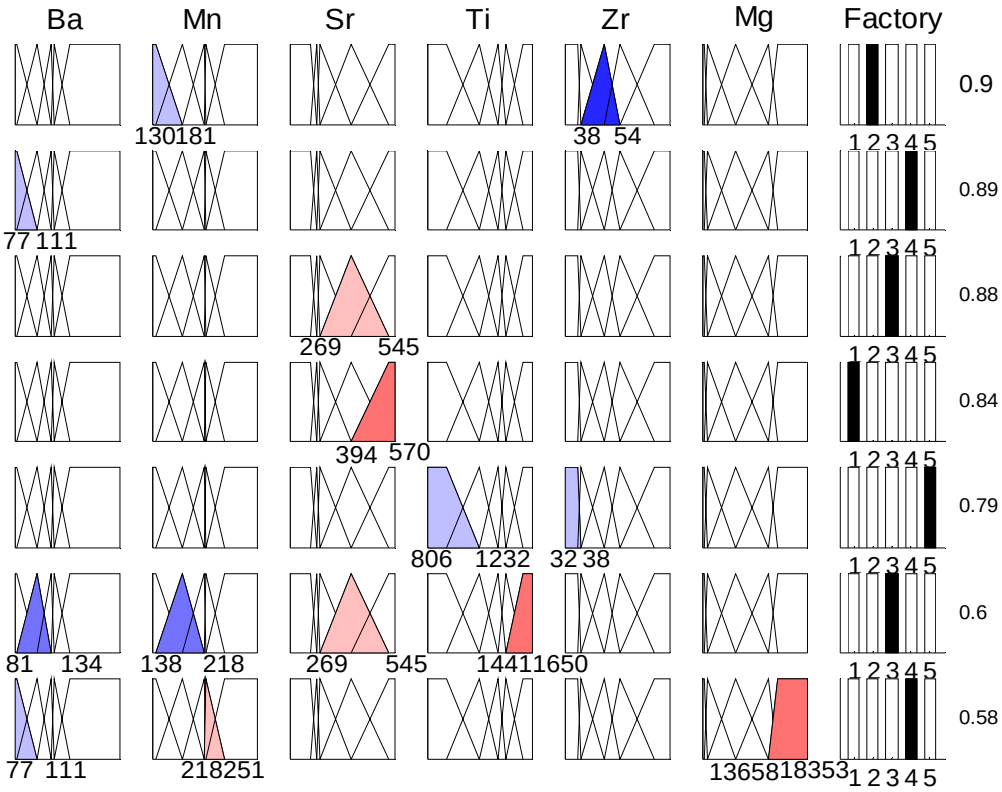


Figure 3 Example fuzzy rule base generated by supervised clustering based approach

| Factory | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th |
|---|---|---|---|---|---|---|---|---|
| 1 | Sr | **Zn** | Zr | Ti | **V** | Mg | Mn | Ba |
| 2 | Mn | **V** | Mg | Zr | **Zn** | Ti | Sr | Ba |
| 3 | Sr | **Zn** | Ti | Mn | Zr | Ba | **V** | Mg |
| 4 | Ba | Ti | Mn | Sr | **Zn** | **V** | Mg | Zr |
| 5 | Ti | Mg | **V** | Mn | **Zn** | Zr | Sr | Ba |

Table 5 Relationship between the factories and trace elements (identification is based on full data set)

| Factory | 1st | 2nd | 3rd | 4th | 5th | 6th |
|---|---|---|---|---|---|---|
| 1 | Sr | Ti | Zr | Mg | Mn | Ba |
| 2 | Mn | Mg | Zr | Ti | Sr | Ba |
| 3 | Sr | Ti | Mn | Ba | Mg | Zr |
| 4 | Ba | Mg | Sr | Mn | Zr | Ti |
| 5 | Ti | Sr | Zr | Mg | Mn | Ba |

Table 6 Relationship between the factories and trace element (identification is based on data set without Zn and V)

The aggregation is applied in both type of identification. Results of supervised clustering based approach are listed in Table 5 and Table 6. Where identification is based on full data set (Table 5) the elements Zn and V are denoted by bold face. The results of aggregation show that the most important trace elements during the determination of origin of clinkers are the Sr, Ti, Ba and Mn. This information is to stand out from the example rule base represented in Fig. 3.

## 5. Conclusions

During the determination process of the origin of clinkers it can happen that the obtained rule-based expert system is unnecessarily too complex as not all the trace elements are needed to identification task. The proposed association rule based method can serve compact and accurate fuzzy classifiers which are easy to use and interpret for engineers and researchers to the analysis of trace elements in clinkers. The method has been implemented in MATLAB© programming language and it will be available from our website: http://www.fmt.vein.hu/softcomp.

## 6. References

[1] R.L. Goguel, D.A. StJohn, Chemical identification of Portland cements in New Zealand concretes, Part I. Characteristic differences among New Zealand cements in minor and trace element chemistry. Cem Concr Res 23 (1) (1993) 59-68; Part II. The Ca-Sr-Mn plot in cement identification and the effect of aggregates. Cem Concr Res 23 (2) (1993) 283-293

[2] J.C. Miller, J.N. Miller, Statistics for analytical chemistry; chapter 7.13: Pattern recognition, Ellis Horwood Ltd., New York, 1984.

[3] F.D. Tamas, Pattern recognition methods for the qualitative identification of Hungarian clinkers. World Cement / Res. & Development 27 (1996) 75-79

[4] F.D. Tamás, É. Kristóf-Makó, Chemical „fingerprints” in Portland cement clinkers in: A. Gerdes (Ed.), Advances in Building Materials Science – Festschrift Wittmann, Aedificatio Publishers, Freiburg – Unterengstringen, 1996, pp. 217-228

[5] F.D. Tamás, A. Tagnit-Hamou, J. Tritthart, Trace elements in clinker and their use as „fingerprints” to facilitate their qualitative identification. In: M. Cohen, S. Mindess, J. Skalny (Eds.), Materials Science of Concrete – The Sidney Diamond Symposium, Honolulu, HI, September 1998. American Ceramic Society, Westerville OH, pp. 57-69

[6] F.D. Tamás, J. Abonyi, Trace elements in clinkers – I. A graphical representation. Cem Concr Res, 32 (8) (2002) 1319-1323

[7] F.D. Tamás, J. Abonyi, Trace elements in clinkers – II. Qualitative identification by fuzzy clustering. Cem Concr Res, 32 (8) (2002) 1325-1330

[8] Dennis H. Rouvray (Editor), Fuzzy Logic in Chemistry, Academic Press, 1997

[9] G. Barkó, J. Abonyi, J. Hlavay, Application of Fuzzy Clustering and Piezoelectric Chemical Sensor Array for Investigation on Organic Compounds, Analitica Chimica Acta, 398 (2-3), 219-22, 1999

[10] F. D. Tamás, F. P. Pach, J. Abonyi, A. M. Esteves, Analysis of trace element in clinker based on supervised clustering and fuzzy decision tree induction, 6th International Congress, Global Construction: Ultimate Concrete Opportunities, Dundee, Scotland, 2005

[11] B. Liu, Y. Ma, C.K. Wong, Improving an association rule based classifier, Principles of Data Mining and Knowledge Discovery (2000), pp. 504-509

[12] X. Yin, J. Han, CPAR: Classification based on predictive association rules, in Proceedings of 2003 SIAM International Conference on Data Mining (SDM'03)

[13] A. Zimmermann, L.D. Raedt, CorClass: Correlated Association Rule Mining for Classification, Discovery Science, 7th International Conference, Padova, Italy, (2004), pp. 60-72

[14] R. Agrawal, R. Srikant, Fast algorithm for mining association rules in large databases, In Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499

[15] J. Han, M. Kimber, Data Mining: concepts and techniques, Chapter 7, Morgan Kaufman, 2000, pp. 279-334

[16] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 7, 773-781

[17] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo,1993