BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 54, No. 4, 2006

Evaluating concrete materials by application of automatic reasoning

D. ALTERMAN* and J. KASPERKIEWICZ

Institute of Fundamental Technological Research, Polish Academy of Sciences, 21 Świętokrzyska St., 00-049 Warsaw, Poland

Abstract. There were two aims of the research. One was to enable more or less automatic confirmation of the known associations – either quantitative or qualitative – between technological data and selected properties of concrete materials. Even more important is the second aim – demonstration of expected possibility of automatic identification of new such relationships, not yet recognized by civil engineers. The relationships are to be obtained by methods of Artificial Intelligence, (AI), and are to be based on actual results from experiments on concrete materials. The reason of applying the AI tools is that in Civil Engineering the real data are typically non perfect, complex, fuzzy, often with missing details, which means that their analysis in a traditional way, by building empirical models, is hardly possible or at least can not be done quickly. The main idea of the proposed approach was to combine application of different AI methods in a one system, aimed at estimation, prediction, design and/or optimization of composite materials. The paradigm of the approach is that the unknown rules concerning the properties of concrete are hidden in experimental results and can be obtained from the analysis of examples. Different AI techniques like artificial neural networks, machine learning and certain techniques related to statistics were applied.

The data for the analysis originated from direct observations and from reports and publications on concrete technology. Among others it has been demonstrated that by combining different AI methods it is possible to improve the quality of the data, (e.g. when encountering outliers and missing values or in clustering problems), so that the whole data processing system will be giving better prediction, (when applying ANNs), or the newly discovered rules will be more effective, (e.g. with descriptions more complete and – at the same time – possibly more consistent, in case of ML algorithms).

Key words: concrete materials, estimation, prediction, design and optimization of the properties of composite materials, Artificial Intelligence, Artificial Neural Networks, (ANNs), Machine Learning, (ML), concrete databases processing, automatic rule generation.

1. Introduction

The possibility of automatic prediction of various properties of concrete is connected with application of new techniques of processing of the data extracted from experimental tests results, especially with reference to the observation of structure of materials. For all those techniques a knowledge base, which is a collection of direct observations of different kind, must first be transferred into a properly formatted database. Such databases can be analyzed by various Artificial Intelligence methods, (AI), called also soft computing techniques, (SC), like artificial neural networks, (ANNs), machine learning, (ML), or certain special techniques generally related to statistics.

Most often the SC methods for extracting knowledge from databases are applied separately, that is one method at a time. The proposed approach is a combined application of several soft-computing methods to estimate, predict, design and/or optimize various properties of civil engineering composite materials, like for example – concrete.

A number of new concepts have been tried in the investigations, like principal components analysis, correspondence analysis, cluster analysis, and some special techniques like rough sets, fuzzy sets, evolutionary algorithms, etc., not always with a convincing result, and not all of them are discussed in what follows.

The motivation for studying possibilities of automatic extraction of the knowledge hidden in examples was to enable rapid taking into account the steady progress in Civil Engineering materials and technologies. Not to be disregarded is a circumstance that the available experimental results that should be taken into account are often incomplete, uncertain, etc. For a human expert even such partial information is of value and so, the automatic system should also make possible its effective exploitation.

Experiments on the proposed approach brought a number of conclusions of practical importance, concerning mainly the prediction of strength and of frost resistance of concrete materials.

2. Automatic reasoning issue

Automatic reasoning is a large and complex area concerning semantic constructions, computational semantics, and so forth. In what follows the issue is simplified to practical techniques enabling computerized inferences concerning experimental observations. Such inferences are needed to uncover as much as possible of the useful information about the technological processes, for which various physical observations have been collected.

Steadily increasing amount of information is available in present technology problems, and such information were previously analysed in a conventional way, mainly by statistical methods. Also, the information was traditionally analysed taking into account only a limited number of selected factors, which, of course, would make building theories and creating

*e-mail: dalter@wp.eu

models much easier but which often is – for various reasons – misleading or impossible.

It is obvious, that the amount of available information is rapidly increasing, both in volume and in complexity. Number of attributes of importance are not any longer a few magnitudes, (like – for example – the amounts of basic components of a concrete mix), but grow up into tens, if not into hundreds of meaningful attributes. Consequently, the only rational and promising way of exploitation of the information is in making its processing possibly automatic. Automatic reasoning involves inferences obtained from application of special computational techniques, mostly associated with artificial intelligence methods, which are implemented as special algorithms of machine learning, various artificial neural networks systems and certain data mining procedures.

An important consideration in automatic reasoning is to delineate the class of problems required to be solved. The analyst must shape in appropriate way the domain of the analysis, which can be very large and should be ordered in some way. Another problem is to correctly indicate various qualitative descriptors, which are carrying information relevant for the process under consideration. Sometimes either the descriptors must be replaced by quantitative data attributes or the whole records must be neglected. Therefore an important question is also what to do with missing or incomplete data, (incomplete records).

In general automatic reasoning concerns not only analysis of datasets which are understood as purely numerical or alphanumeric matrices, but it concerns also special problems, such as identification and recognition of images or recognition of speech, making proofs in propositional logic or in predicate logic, playing games, etc. These possibilities will not be discussed here.

There is a difference in the automatic reasoning if compared to traditional expert systems, based on the idea of an expert, who should be able to predict every possible result of analysis, thanks to certain previous knowledge. An automatic reasoning system should be able to create also new answers, i.e. such results, that nobody did forecast before. Automatic reasoning is, however, not dreams or imaginations, but a pertinent reasoning over the available facts.

An important addition to automatic reasoning techniques are genetic algorithms and similar solutions, which allow analysis of certain numerically very complex problems, for example in optimization.

In the present approach selected from the multiple possibilities of automatic reasoning are only two most important possibilities: artificial neural networks, (ANNs), to predict numerical values, and machine learning algorithms, (ML), to identify relations, associations, etc. These are or may be supported by techniques originating in statistics, making possible the evaluation which ones of the attributes describing the given process are important, or in which order the records and attributes should be set to facilitate a more efficient analysis of the problem.

3. Preparation of experimental database

The data for experimenting with the proposed approach were collected from available reports or other publications, and from special laboratory experiments at IFTR, [1]. Although there is plenty of information reported on behavior of concrete materials, there is no standard concerning how the data should be formatted and described for civil engineering purposes. The issue is seldom treated, and one of rare examples in this field is the paper published by the American Concrete Institute, [2].

The most typical data in materials technology are quantitative and these are mainly the direct results from various physical or chemical measurements. Together with the accompanying descriptions and comments the observations create a knowledge base, (KB). A knowledge base may contain also reports, photographs, audio recordings, etc. The proposed approach concerns only a certain subset of those. The results selected from KB and formatted in appropriate way generate a database, (DB), which is a kind of table, (generally speaking: an alphanumeric table), built of records, with each record composed of attributes. The attributes can be rational numbers, nominal descriptors, etc., (other possible attributes types, not discussed here are, for example: logical, cyclic, continuous, [3]).

Databases in which records are only numbers can be processed by statistical programs and by Artificial Neural Networks, (ANNs). Basically, in such case analyzed can be databases with only complete records. This means that each record must have all the attribute values specified, and there should be no lacking values in the whole database. Such ideal situations are often exceptional in engineering practice.

In a real database, however, certain records may be defective and/or have blank attribute values. Such records must therefore be either reconstructed or excluded from further processing. A common approach, (not always effective), is that the necessary reconstruction of the data can be done by analysis of the remaining, complete records from the same database.

Apart of the problems with uncertain and defective data there are difficulties connected with processing of the qualitative data. This happens often in Civil Engineering reports that the observations cannot any longer be presented as numbers, (scalars, vectors, etc.), but rather as symbols – simple or more complex. Such symbols can characterize the proper names of the raw materials or the names of producers, geographical names, descriptors given by experts, descriptors of colors, odors, etc. Trying to work on a database with these types of variables, when applying almost any type of ANNs, (there are a few exceptions), the qualitative attributes could be coded into numbers, but for various reasons such a simple solution is rather risky.

A possible, although computationally complicated solution in the above mentioned case is introducing new attributes, of Boolean type, systematically replacing one by one all the legal values of each nominal attribute present in the database. In such case all new attributes will have the value of 1, (one), if the corresponding state is “on”, or 0 (zero) in the opposite case, (the state: “off”). Naturally, the resulting new database will have much larger volume, and needed will be an additional special tool for database processing, (for its coding and de-coding).

Typically, a technological observations database will be a collection of different components. An example of a database on concrete mixes can be seen in Table 1, and its structure is explained in Table 2. In the Table 1 the meaning of the special sign '?' is that there is a lack of any information at all, concerning the respective value of the attribute.

In the above example the observed features concern composition and some hardened material structural properties of ordinary concrete or High Performance Concrete, (HPC), related to their compressive strength and their frost resistance. There were 42 attributes and more than 1700 records in the database, [1]. Some data for testing of the procedures were taken also from virtual databases, prepared by simulation.

A database as in Table 1 can usually be treated as a general database, which will be a source for selection of its certain subsets: certain "working" databases. To characterize and select such sub-base will only be possible after the aims and the scope of the whole data processing is decided by the operator, (by the user of the system). There is no accuracy defined for the particular attributes of such database, as the data are usually obtained from various external sources, (papers, reports, any kind of documentation), and their accuracy can never be better than as they were submitted.

An important element of the database preparation may be generation and selection of the so called derived variables, which are functionally related in some way to the original components, (typically by algebraic operations on particular attributes). For example the water-cement factor will be such a derived attribute, made of the primary attributes of cement content, and of water content in the concrete mix.

4. The applied AI tools

A number of soft computing tools were tried in the present investigations, for example: [5–12]. The algorithms selected finally to process the data in the proposed system were those of BP ANN, Fuzzy ARTMAP and aiNet artificial neural networks, aq19 and See5 machine learning programs, the GradeStat, and certain statistical procedures available in SPSS or in Statistics Toolbox of Matlab, applied for clustering of the data.

The ANNs solutions are relatively well known and often used in engineering, but popularity of the other algorithms is limited in problems of Civil Engineering materials. This concerns especially Machine Learning (ML) programs – that is searching of the rules by a sequential covering or in form of decision trees.

Table 1
Fragments of a database on the properties of concrete

No	...	FA_L [kg/m^3]	SF [–]	$density$ [kg/m^3]	air [%]	$alpha$ [1/mm]	L [mm]	A_T [–]	...	fc [MPa]	$Sc56$ [kg/m^2]	$Sc112$ [kg/m^2]	DF [–]
1	...	0	0	2475	2.21	22.96	0.30	gravel	...	84.6	0.216	0.75	?
2	...	0	0	2498	4.05	13.01	0.37	gravel	...	91.7	0.153	0.70	?
3	...	99	0	2372	4.00	20.63	0.24	gravel	...	79.3	0.178	0.69	?
4	...	0	1	2466	4.92	21.77	0.23	granite	...	73.9	0.070	0.26	?
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1705	...	0	0	?	4.10	25.60	0.25	?	...	?	?	?	95
1706	...	0	0	?	4.60	28.00	0.16	?	...	?	?	?	98

Table 2
Definition of the structure of the database as in Table 1

Symbol	Attributes	Units	Type	Min	Max	Attributes if nominal: No of cases	Attributes if nominal: List of values	Remarks
...	...	...	...	...	...	...	...	...
FA_L	Lightweight aggregate presence	[kg/m^3]	con	0	1035			
SF	Silica Fume presence	[kg/m^3]	con	0	298			
$density$	Density	[kg/m^3]	con	1340	2504			
air	Total air content	[%]	con	2.21	7.40			from IA*
$alpha$	Air voids specific surface	[1/mm]	con	4.04	44.00			from IA
L	Air voids spacing factor	[mm]	con	0.12	1.15			from IA
A_T	Aggregate type	[–]	nom			3	gravel, granite, basalt	
...	...	...	...	...	...	...	...	
fc	Strength 28d	[MPa]	con	23.30	99.50			after 28 days
$Sc56$	Frost resistance. Borås method, (1)	[kg/m^2]	con	0	1.5			56 cycles
$Sc112$	Frost resistance. Borås method, (2)	[kg/m^2]	con	0	2.2			112 cycles
DF	Frost resistance ASTM C666	[–]	con	0	110			DF*

* IA – image analysis, DF – durability factor, [4]

Both ANNs and ML solutions can be supported by other algorithms like those originating from statistics or from evolutionary computing, (to select optimal network architecture), but these possibilities have not been employed in the proposed system. For various reasons not applied were also the concepts from PCA, (Principal Components Analysis) and those of the rough sets approach.

Artificial Neural Networks, (ANNs) are a promising tool in case of purely numerical datasets. There is a number of possible solutions available – e.g. [6,11–13]. The most popular are different feed-forward, back propagation networks, build of external and hidden layers of neurons with appropriate, adaptable weights. A simple network of this type was created also in [1], (as BP ANN).

However, the best choice among various concepts concerning architectures of ANNs was found to be the one originating from ART (Adaptive Resonance Theory), called Fuzzy ARTMAP, [14]. Its algorithm enables effective proximity evaluation in multidimensional space of selected attributes of purely numerical records. It happens that this solution does not always realize the prediction task. Sometimes it refuses response, which seems to be a disadvantage, but as a matter of fact is a very positive feature: when the network is not properly trained, due to lack of appropriate data, (similar to the data to be predicted), the network should refrain from any prediction at all, rather then suggesting some numbers, which may be completely unrealistic. This positive feature is absent in many other ANNs solutions.

In the experiments sporadically was applied also a pseudo-ANNs procedure of aiNet, [6], but this was used only for quick checking certain solutions of ANNs.

All the ANNs systems that were tried, perhaps except IAC, [13], require purely numerical data. The special case of the IAC network, (Interactive Activation and Competition), does not only make possible to analyze data described with nominal attributes, but also to work on incomplete data. This solution, however, seemed little practical for the present system.

In case of Machine Learning programs, (ML approach), the analyzed data are treated in terms of classes and relations among them, not only in the terms of numerical values. Because of that, the qualitative descriptors which introduce difficulties in case of most ANNs solutions are quite natural here. The system can automatically construct rules for classes defined by opposite phenomena, involving relations on selected attributes, for example for the groups of frost resistant and non frost resistance concretes.

In the example below – Fig. 1 – the system is also to identify rules, (usually more than one rule), for discrimination of the submitted records as belonging to different classes. The property in question is in the present case the resistance of concrete to freezing and thawing actions. Among others it depends on density of concrete, (density), air contents in the hardened concrete, (air), so called Spacing Factor, which characterizes spatial distribution of the air bubbles generated by the air entrainment, (L), and so forth, (cf. Table 2).

Analyzed are databases for two classes of examples, described by various combinations of the attributes, (density, air, L, etc.) which must be properly defined previously, (like in Tables 1 and 2). The attributes are from actual observations, so unavoidable is a varied accuracy of the input data, (like in the case of the above attributes L and $Sc56$).

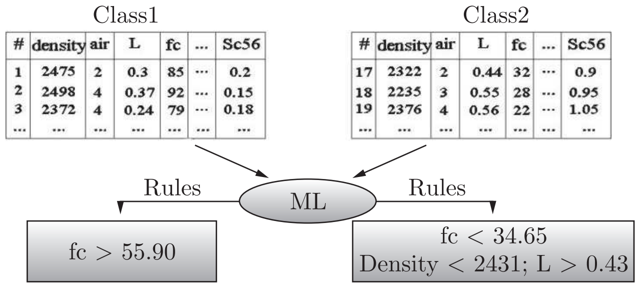

Fig. 1. Idea of generation of rules by ML program

The quality of the hardened concrete is characterized by its compressive strength, (fc), and by the amount of material detached from its surface during 56 cycles of freezing-and-thawing, ($Sc56$). No one of such parameters describes completely the property in question, as hardened concrete can be as well strong and NON-frost resistant as opposite: weak, but of a good resistance to the action of frost.

In the present experiments applied were two ML solutions: aq19, [7] and See5, [8]. When the programs learn general decision rules from examples gathered in decision classes, the decision rules are optimized according to certain user-defined optimality criteria. For example the aq19 family can refer to syntactic simplicity of rules, (measured by the number of proposed hypotheses), to number of conditions in the rules, the simplicity of the conditions, or to a combination of all these factors, plus a certain evaluation of the cost of creation of the rules, through the cost of acquiring the attributes involved in the rules. The processing goes like in the symbolic example in Fig. 1. For both classes of examples an ML system generates separate rules – often called hypotheses. Examples can be seen at the bottom area of Fig. 1.

Among the records taken into consideration those of class 1 represented frost resistant mixes, those in class 2 – mixes of poor frost resistance.

The pre-processing of Civil Engineering data involves also identification of the outliers. It should be followed by their evaluation, because the apparent outliers may not necessarily be a false information.

A very important element of the data processing is finally clustering and reordering of the records. Such action appeared especially promising in case of GradeStat algorithm [9,15], where for the re-ordering of the database applied is the concept of a so called correspondence map.

In all the experiments positive results were obtained by introducing derived attributes. Only in very rare cases it was found profitable to replace the nominal or structural attributes by continuous attributes.

5. Proposed system

The three kinds of the applied tools were artificial neural networks, (ANNs), machine learning programs, (ML) and the algorithms related to statistical data analysis. The typical datasets were limited to those that can be represented by series of numbers and symbols, in form of alphanumeric strings.

The proposed system was aimed at analysis of data from civil engineering domain, but much of it can be utilized also in many other engineering problems. Instead of technological descriptions, like those concerning concrete composition, its various physical features, etc., the system can therefore analyze other materials, various technological processes, also instances of building failures, reasons of deterioration, subjective descriptions of the quality of the object as estimated by experts, etc.

A simplified layout of the proposed system for data processing is presented in Fig. 2.

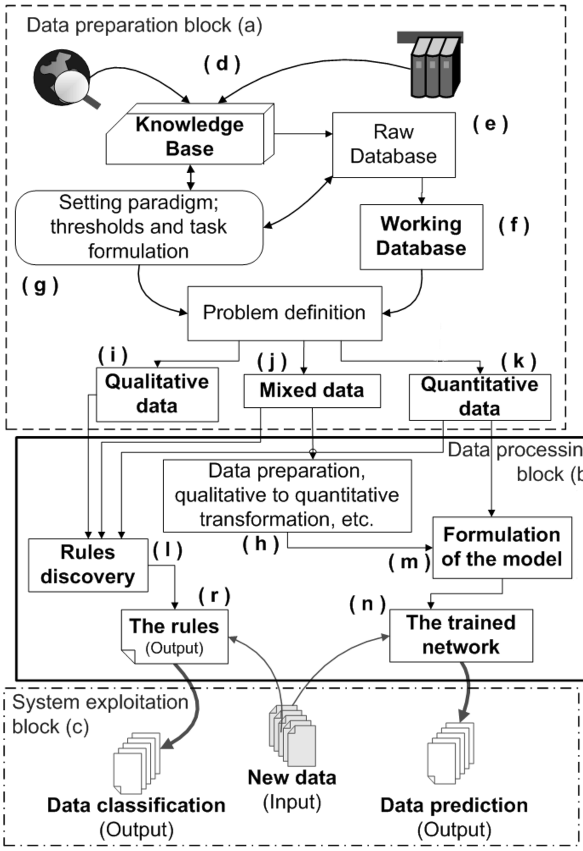

Fig. 2. The main parts of the proposed system

The system is built of three main blocks, dedicated to data preparation, data processing and training, and to system exploitation. The whole system is complex and too complicated to show it in a single drawing. As an example in Fig. 3 presented is only a small sub-block to explain certain details of the scheme, [1].

A typical procedure starts with identification of the defects in the dataset and their repair. Then the data are reordered, divided and clustered, taking into account possibilities of GradeStat and other statistical tools. Finally the results are used for training and prediction, or for rules generation, depending on the kind of data under consideration, (qualitative, quantitative, or mixed).

The idea was to apply different AI tools combined, and to enable automatic processing of huge datasets. In this way the system and its components represent a hybrid approach to data recognition and data analysis. It can be seen in Fig. 3 that choice among different options is possible; the user must in such case decide which one of several solutions is to be applied. For example – the upper, right hand area in the Fig. 3 – the user in certain step can select among correspondence analysis (GradeStat), cluster analysis or some heuristic solutions.

In the exploitation block of the system there are mainly two actions expected, which are data prediction and generation of rules, (bottom block of Fig.2). Which one of the two above actions is applied is also an optional decision of the user but – of course – the data can also be collected and processed without any particular task definition. In such way only the data preparation block will be used, to enable a further analysis, or simply to store the experimental data for future exploitation

If the data are defective, and a number of values of some numerical attributes are lacking, then each lacking value can be predicted using the ANNs trained with data from the remaining set of complete records. Of course there are natural limits of the procedure related to the proportion in the dataset of defective records. In [1] discussed were other methods of predicting unknown values of attributes, like basic statistics, empirical formulas, and possibilities of the correspondence analysis.

The clustering allows identification of data subsets for which certain relations can be searched for, as being expected different from those in other subsets, [16]. For the above purposes applied were algorithms like nearest neighbour clustering and the correspondence analysis.

6. Obtained results

An analysis of the collected experimental databases were carried out to improve prediction of compressive strength of concrete by using three artificial networks. This is an alternative to the traditional way of designing concrete mixes. This might also have an important economical aspect, because the procedure may allow designing cheaper compositions of concrete mixes.

The experiments were conducted on databases concerning composition, hardened material structure and properties of concretes. Some data processing experiments concerned also specially prepared virtual databases, similar to the real ones, but with user's full knowledge on the relations among the attributes.

From certain initial experiments a conclusion was drawn that the order in which data are submitted to the analysis may influence the results of predictions. The observation seems to concern both the order of the attributes and the order of records. For this reason the database was first analysed using GradeStat package, which applies statistical grade exploratory methods, based on calculation of correlation matrices and certain measures of concentration.

By evaluating differences between the values of particular attributes, (particular variables), the GradeStat program creates a two dimensional, grey-scales map representing records composed of attributes, (a grey-scales or multiple colours map). The intensity of the colour is selected according to so called "over-representation" factor. The map allows the user recognition of groups of records and groups of attributes that are in certain meaning "similar".

In one of the experiments processed was a database on the compressive strength of HPC concrete. In the database, which was purely numerical, each record was represented by 11 input attributes and there were about 200 records concerning HPC. From the database 150 records were later used as the training set, and the rest – as the testing set. In Fig. 4 displayed is the result of reordering the whole database in the GradeStat, (the over-representation map), which the operator evaluates by a visual inspection.

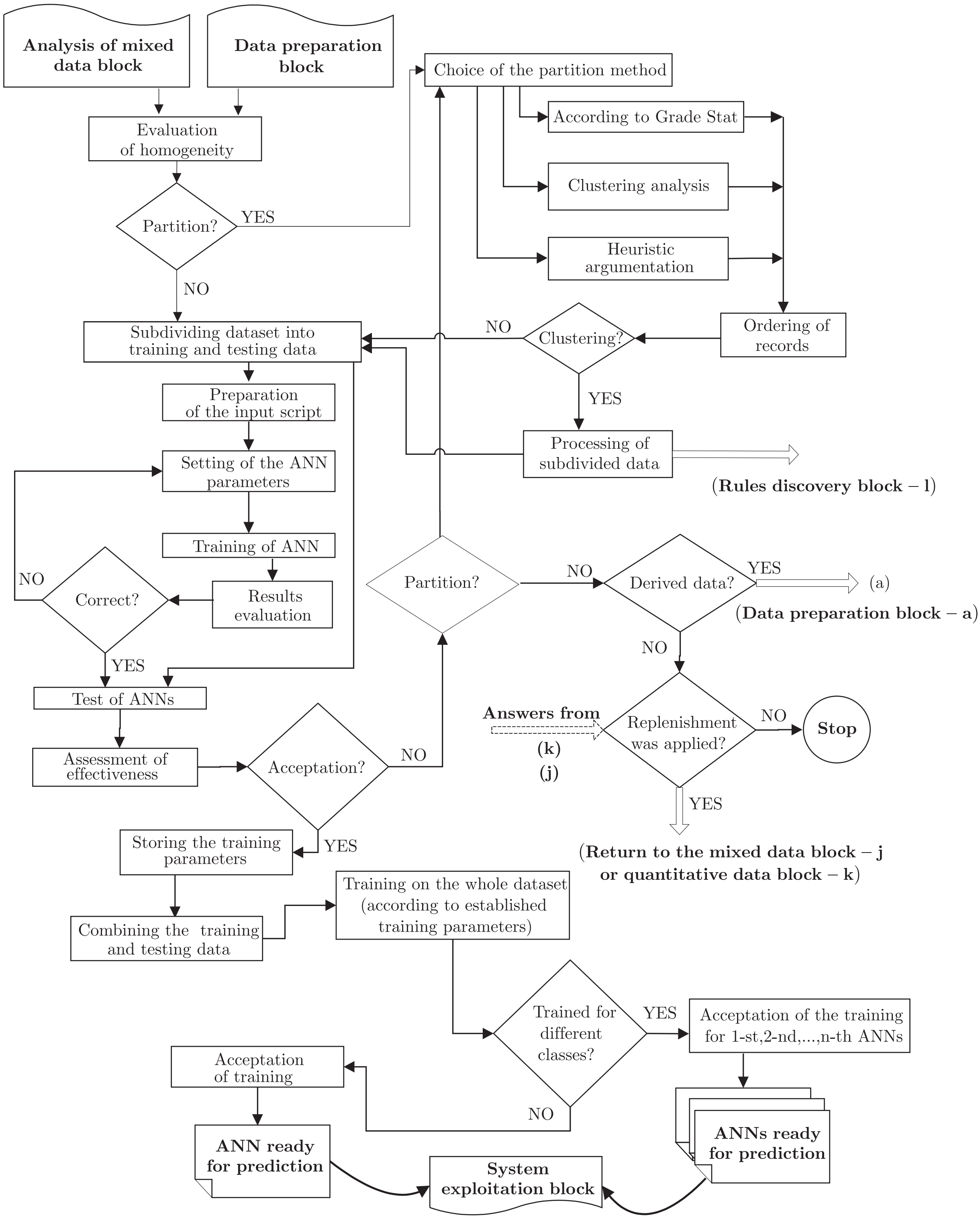

Fig. 3. An example of a component of the system: this is a sub-block from the block (b), Fig. 2, for quantitative data analysis, (block m)

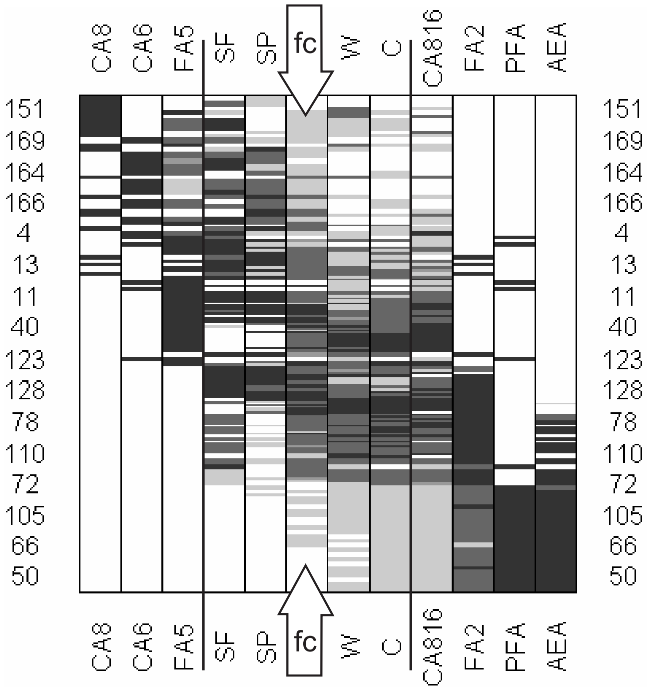

Fig. 4. A correspondence analysis map created by GradeStat for a small dataset – matrix of 200 records with $1 + 11$ variables (attributes). Rows represent the records, and columns – the attributes of the dataset. The structure of the shading is correlated with similarity-based distribution of – simultaneously – records and attributes

In the figure the numbers identify symbolically selected records and the letters identify the attributes: *FA*2 – fraction of fine aggregates up to 2 mm, *FA*5 – fraction from 2 to 5 mm, *CA*6 – coarse aggregates 5÷6 mm, *CA*8 – 6÷8 mm, *CA*816 – 8÷16 mm.

In effect of the analysis certain initial, original order in the database was annihilated and the image reveals new order, which should be taken into account when exploiting the data. It can be seen that the attributes concerning the description of fraction of aggregates, (*CA*8, *CA*6, *FA*5, *FA*2, *CA*816), and additives and admixtures, (*PFA* and *AEA*), have been located far from the main output attribute of the 28 days compressive strength, (*fc*). Only four attributes, concerning contents of silica fume, (*SF*), superplasticiser, (*SP*), water, (*W*), and cement, (*C*) were located close to the output attribute of the compressive strength. It was decided then, to search for predictions taking into account attributes of different degree of closeness to the target attribute of *fc* – i.e. various sets of its “neighbour” attributes.

After re-ordering predictions were obtained for the output attribute (*fc*) applying Fuzzy ARTMAP (one of ANNs solutions), taking into account either all 11 input attributes, or selected 8 attributes, (using as combined the fractions of two categories – fine and coarse aggregates, respectively – FA and CA), or only selected 4 “closest” attributes, (*SF*, *SP*, *W*, *C*). The average prediction errors were diminishing in this test, being – respectively: 19%, 14% and 11%. This demonstrates possible improvement of the predicting power of the system by elimination of some of the attributes from the dataset. A similar effect is expected in case of ML approach to data processing, but in this case the results are not yet fully conclusive.

Another example of the proposed system possibilities concerned evaluation of the frost resistance of hardened concrete. The dataset – another sub-sets of the original database, concerned real data from tests performed according to four different national standards, [4], to evaluate internal destruction of concrete, or the external concrete surface scaling by cyclic changes of temperature.

The experiments were performed applying ML programs to generate rules for frost resistant and non frost resistant concrete materials.

A number of observations have been done in these tests. An important problem, which seems to be typical for investigations similar to those presented in this paper, was a deficiency of proper quality data in electronic form. Even if in civil engineering many records are nowadays stored electronically neither the care of the data is a common practice nor is it done correctly and systematically. Usually data mining demonstrations are therefore possible mainly on data that are of questionable quality. In this respect the situation seems to be different from that in e.g. medical sciences, where the value of records on patients and on their illnesses is generally well recognized.

A typical search for a satisfactory rule that both has relatively large support and is consistent (free of contradictions) involves repeating trials of application of different combinations of the control parameters. What is obtained are tables of rules, usually ordered according to decreasing value of factor *q*, denoting the rule quality criterion, (based on rule coverage and training accuracy). Usually the first rules, presented at the beginning of the output list and expected to be “the best”, either cover many positive records of the database in question, and at the same time many negative examples, or - the opposite - they result in a low number of negative examples, but also low number of positive coverage. An example of such a table, taken from another, much simpler investigation, is shown in the first text box at the top of the next page.

In this output the selectors are marked by square brackets. The four letters symbols, like LzdM, sazM, etc., represent certain linear type variables. The resulting conjunctions of selectors – the rules – are characterized by descriptors enhanced in the text box by the bold italic, (they can be seen on the right hand side of the textbox). The user of the AQ19 program is searching among multitude of similar results, looking for a satisfactory combination of parameters t, (t specifies the total number of positive examples covered by the rule; a.k.a. rule coverage or its support), and n, (n denotes the number of negative examples covered by the rule; a.k.a. negative coverage, or exceptions, or conflicts).

The rules are produced with different number of conditions and to get a generalization effect the user is usually interested in simple rules, that is in the rules with a possibly low number of conditions. From such a point of view the most “interesting” would be for example the last rule in the frame, (Rule No 9), really simple, with only two selectors, which on the other hand, is completely useless, as it properly describes only a single record among hundreds, (!). So, the proper selection of good rules needs experience of the operator.

```
aggregate-outhypo
    #   rule
    1   [LzdM=1..175,396..471] [sazM=11..13,17..50]
        [sazH==0..121] [LzdH=0..13,44..158]
        [sazL=11..19,22..283]
        [senH=0..382] [senM=22..65,375..815]
                                           (t:125, u:114, f:2, n:54, q:0.422617)
    2   [LzdM=23..496] [sazM=0..114]
        [sazH=61..176] [senL=0..430,665..1075]
                                           (t:42, u:39, f:1, n:9, q:0.311317)
    3   [LzdM=9..22] [sazM=28..35]         (t:13, u:12, f:0, n:0, q:0.214707)
    4   [LzdM=8..334] [sazM=15..32,72]
        [LzdH=83..164]
                                           (t:13, u:10, f:0, n:2, q:0.184212)
    5   [LzdM=1..2]                        (t:9, u:4, f:0, n:0, q:0.178647)
    6   [LzdM=3..16] [sazM=18..27] [sazL=58..125]  (t:4, u:4, f:0, n:0, q:0.119098)
    7   [sazM=33..38] [senL=638..1390]
        [sazL=44..65] [LzdL=51..71]
                                           (t:4, u:3, f:0, n:0, q:0.119098)
    8   [LzdM=253..342] [sazM=30..42]
        [senL=268..292]
                                           (t:2, u:2, f:0, n:0, q:0.0842152)
    9   [sazM=37] [senH=775]               (t:1, u:1, f:0, n:0, q:0.0595491)
```

```
parameters
run mode echo verbose run ambig trim wts maxstar criteria noise q_weight neg_ex_probe
1   ic   pv   1       1   neg   mini cpx  1       default  yes   0.50     0
```

It was observed also that the whole issue of finding an appropriate rule, or a good "hypothesis" in concrete materials design, is an extremely nonlinear problem. Experiments on certain civil engineering data demonstrated that the results of the experiments may be such that statistics of the rules created by the system are not related systematically to the system parameters, and there is a multitude of such parameters. For example, in the case of AQ19 some parameters are: *mode, verbose, ambig, trim, maxstar, noise, q_weight, neg_ex_probe*, not to mention eleven pre-defined criteria that can be set in various combinations, with different tolerances and at a different – so called – cost of variables.

An example of some of the AQ19 parameters can be seen in the second frame above.

The meaning of such parameters, which were applied to obtain rules (a) and (b) presented in what follows, is e.g.:

— trim=mini – the program is to create rules as simple as possible, each with a minimum number of conditions, and with a minimum of values,
— noise=yes – the training data is assumed as potentially containing "noise"; the alternative - noise=no - would mean the dataset is an infallible oracle, with no errors,
— ambigue=neg – when rules are created the ambiguous examples should always be taken as negative examples,

(in this second frame the variables are described in the upper row, their values, respectively, in the row below).

The additional observation is that the approach to such a strongly nonlinear problem can be either done by experimenting, or the system should be furnished with a tool for some automatic selection of those solutions which are expected to be "better". It seems that the creators of AQ algorithms started subsequently to work also on such approach, but within the proposed system the search for "better" solutions was done manually. Because of the nonlinearity issue the system described in this paper can not work in a fully automatic way, and it needs frequent interventions of the user. This concerns all the methods involved: ANNs, ML, and various statistical methods.

Certain additional ML experiments were performed recently on a database of 541 records, each described by 42 attributes, but very imperfect – containing on average 63% of lacking (unknown) values, with the symbol '?' in the database. These experiments were also giving realistic answers, although in similar case the resulting rules can be expected to be of relatively small support.

The system produced finally simple results as follows. The concrete is expected to be frost resistant if condition (a) is realized. The concrete will be non frost resistant, when condition (b) is true:

$$[w\_c < 0.43]\ [fc > 55.00] \tag{a}$$

$$[w\_c > 0.32]\ [A\_hr < 3.6]\ [fc < 57] \tag{b}$$

The meaning of the symbols in these equations are: w_c – a water cement-ratio, (the simplest example of a derived variable), A_hr – volume of the air observed in the hardened concrete evaluated by IA, (image analysis), and *fc* – the compressive strength.

The above rules were confirmed with experiments performed at IFTR, for frost and non frost resistant concretes estimated by the Borås method, [4]; in this case obtained was 100% correct classification.

7. Concluding remarks

The proposed approach was verified with results from experiments performed in the past years at the IFTR laboratory. Good correlations were observed also when operating on simulated (virtual) databases.

It should be added, however, that even if the similar procedure should generate correct rules, they will not extrapolate the knowledge beyond the original source database domain. By no means such system could ever be considered as a fortuneteller, (!).

Employment of various AI tools makes possible to improve quality of the data, (by elimination of outliers and/or missing values, or by applied clustering and re-ordering of the data, etc.) Data re-ordering seems to be helpful to improve the effectiveness of ML procedures.

By applying the proposed system it was possible to evaluate the frost resistance quality of hardened concrete, independent of which frost resistance measurement method was applied in its examination. Certain generalized rules were obtained, and this is a positive, to a certain degree even unexpected result, not suggested before in concrete technology.

The present results concerned the field of concrete technology, but they could also be important for application in various other domains of civil engineering, probably also in structural analysis problems.

Acknowledgements. The paper is based on a dissertation [1]. Its preparation was supported by a grant No.: 5 T07E 059 24, from the Polish Committee for Science and Informatics, (MNiI). For the ML experiments the AQ19 program was obtained from MLI Laboratory, George Mason University.

REFERENCES

[1] D. Alterman, *Evaluation of Concrete Materials by Automatic Reasoning*, Doctoral Dissertation, IFTR PAS, Warszawa, 2005, (in Polish).

[2] C.B. Oland and C.F. Ferraris, *Guide to Recommended Format for Concrete in Materials Property Database*, ACI Committee, 1999.

[3] R.S. Michalski and K.A. Kaufman, *The AQ19 System for Machine Learning and Patterndiscovery: a General Description and User's Guide*, George Mason University, 2001.

[4] A.M. Brandt and J. Kasperkiewicz, *Diagnosis of Concretes and High Performance Concrete by Structural Analysis*, IFTR PAS, Warszawa, 2003, (in Polish).

[5] SPSS for Windows, `http://www.spss.com/`.

[6] *aiNet – ANNs*; a trial version, `http://www.ainet-sp.si/NN/En/nn.htm`

[7] *AQ19 – Machine Learning Programs*, `http://www.mli.gmu.edu/mdirections.html`

[8] *See5/C5.0 – Demonstration*, `http://www.rulequest.com/download.html`

[9] *Gradestat – Grade Correspondence Analysis Package*, `http://www.ipipan.waw.pl/ gradestat/`

[10] *ROSETTA – a Rough Set Toolkit for Analysis of Data*, `http://rosetta.lcb.uu.se/general/download`

[11] *Statistics and Neural Network Toolboxes of MATLAB*, `http://www.mathworks.com/web_downloads`

[12] *Stuttgart Neural Network Simulator*, `http://www-ra.informatik.uni-tuebingen.de /SNNS`

[13] *IAC – Interactive Activation and Competition*, `http://www2.psy.uq.edu.au/ brainwav/ Manual/IAC.html`

[14] J. Kasperkiewicz, J. Racz, and A. Dubrawski, "HPC strength prediction using artificial neural network", *Journal of Computing in Civil Engineering* 9 (4), 279–284 (1995).

[15] T. Kowalczyk, E. Pleszczyńska, and F. Ruland, *Grade Models and Methods for Data Analysis with Applications for the Analysis of Data Populations*, Springer-Verlag, Berlin, 2004.

[16] J. Kasperkiewicz and D. Alterman, "On effectiveness of preprocessing by clustering in prediction of CE technological data with ANNS", *Int. Conf.: New Trends in Intelligent Information Processing and Web Mining*, 261–266 (2003).