ON THE IMPACT OF MACHINE LEARNING APPLIED ON RESEARCH CHARACTERISATION DATA; A NANOINDENTATION CASE STUDY

K. Paraskevoudis[1,2], E.P. Koumoulos[1,2,*], C. Charitidis[2]

[1]Innovation In Research & Engineering Solutions (IRES), Boulevard Edmond Machtens 79/22 - 1080 Brussels, Belgium

[2] National Technical University of Athens, School of Chemical Engineering, RNANO Lab – "Research Unit of Advanced, Composite, Nano Materials & Nanotechnology", 9 Heroon Polytechneiou Street, GR-15773, Zographos Athens, Greece

(*elikoum@chemeng.ntua.gr)

ABSTRACT

Applications from the characterisation field produce huge amounts of Data, which can be further exploited with Machine Learning. Data driven algorithms can help study big Data and reveal underlying structures and hidden patterns of Characterisation Data. In the present study, Nanoindentation Data from mortar grids were analysed with K-Means algorithm and Random Forest Classifier in order to approximate the initial surface constituent phases of the material and reconstruct its topology. One 49-points dataset (7x7 mortar grid) with Modulus (E) and Hardness (H) measurements was contained from Nanoindentation experiments and further studied with K-Means algorithm, in order to evaluate the approximation of Machine Learning in mapping the above points to the various constituent phases of the mortar. Resulted clusters were cross-validated with bibliography values. In addition, a Random Forest classifier was trained under 49-points Dataset with E and H measurements and the corresponding surface colour of the mortar grid (obtained from optical microscope) in order to predict the surface colour from E and H values. The Random Forest model was then used to make colour predictions on the first Dataset. The predicted colours were studied as of possible correlations with the resulted clusters (constituent phases) from K-Means and it was shown, that each cluster corresponds to a unique surface colour. Furthermore, the constituent phases of the first mortar grid was reconstructed as of the resulted K-Means clusters.

INTRODUCTION

Experimental Data coming from the Characterisation Field often stays unexploitable besides its initial purpose of classifying the material or some of its properties. More often, this Data comes with great complexity making it even harder for Field Experts to analyze it. However, the huge spreading of Machine Learning offers new opportunities and can help exploit generated characterisation Data and overcome almost any difficulty. Underlying structures and hidden patterns of Characterisation Data, which are not obvious or readable by humans, can be revealed with computational applications[1,2].

Machine Learning is a growing field of application in Materials Science especially for structure-property relationship prediction, crystal structure prediction, micrograph analysis and descriptors identification[2,3,4]. Furthermore, data driven algorithms have helped automate the determination of phase diagrams using high-throughput combinatorial experiments[3].

Machine Learning approaches also contribute in advanced and smart materials modelling, where several algorithms are used for reverse engineering. In this case, regressions and classification models can help predict the design parameters of a material (optimization) in order to satisfy

some desired properties[5]. Regression algorithms can be also used for the prediction of mechanical properties of constructive materials[6].

In general, Machine Learning is a cross-section Field of Informatics and Mathematics, which contains various algorithms that learn from prior knowledge and data and then make predictions[7,8]. In most of cases, Machine Learning approaches try to build a function which best models the given Data. This process is called the training procedure of the model[9]. Depending on the existence of a desired label to predict, Machine Learning algorithms are divided into supervised and unsupervised learning. In supervised Machine Learning, the goal is to model the input variable (x) respectively to a label (Y) and learn a function Y=f(x) with the least error. The input variable (x) are values of several attributes that may influence and weight in the label (Y). The label (Y) can be a continuous value (regression) or a category (classification). In both cases, the model is trained under labeled Data. In regression, the trained model aims to predict a real value whereas in classification the task is to predict a class or category, which has no numerical significance. In the other hand, in unsupervised learning there are no labels. Unsupervised algorithms try to best group (cluster) Data based on mathematical similarities of the attributes.

K-Means is a typical and popular example of unsupervised learning, which is used for the purposes of this study. It aims to mine through unlabeled Data and to partition all of the Dataset's observations into k clusters. Each observation will belong to only one cluster, which is the one with the lowest metric of distance. The first step of the algorithm is to initialize random k cluster center points. Each observation is then assigned to the cluster with the smallest Euclidean distance to it (or any other selected metric). In the next step, the means of each resulted cluster are calculated and the previous cluster center points are moved to the respective cluster mean. Each observation is re-assigned to the nearest cluster. The previous steps are repeated until the clusters' centers can no further be moved[1], as show in Figure 1.

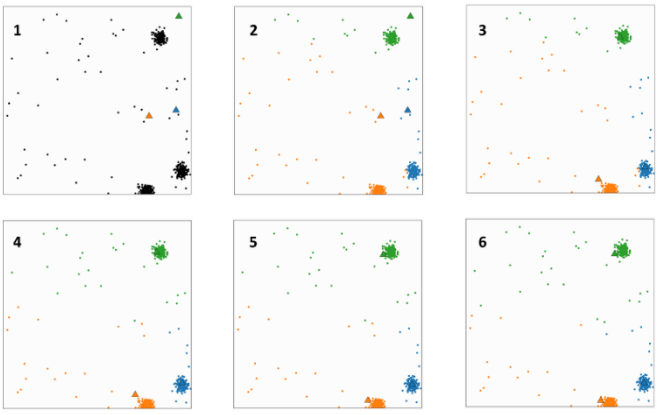

Figure 1. An example of K-Means clustering with 3 Clusters. In the first step are the raw unclustered data, where Three random centers are then assigned. In step 3, the data points are assigned to the center with the smallest Euclidean distance and the corresponding clusters are revealed. This procedure continues in the next steps until the algorithm has converged and the clusters' centers can no longer be updated (step 6).

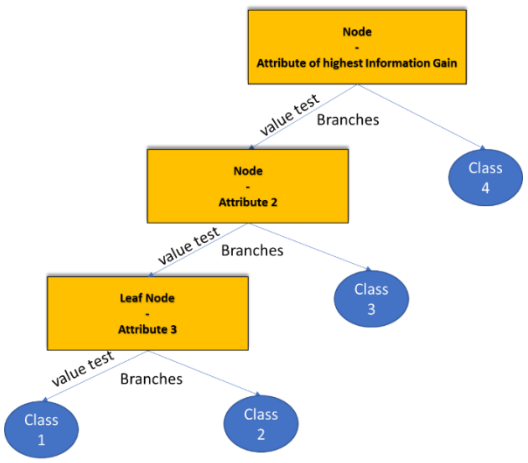

Figure 2. An example of a Decision Tree

Random Forest classifiers are a variation of the classical Machine Learning Decision Trees, which are supervised algorithms that try to model given labelled Data and are then used to make predictions on new instances. The typical Decision Trees consist of nodes, branches and leaves. In each node the feature values are tested. Depending on the value of each test, there is one corresponding branch. All nodes and branches finally lead to the leaves, which signify the class labels (output variable). The biggest challenge in training a decision tree is to select the hierarchy of attribute testing and more specifically to find the root attribute to start splitting on in the decision making. For this purpose, the information gain of each attribute is calculated, which signifies the importance of each attribute. Information gain is the amount of information obtained about a random variable (here our attribute) from observing another random variable (here our label Y)[6]. Attributes with higher information gain are tested first in the decision tree. An overview example of decision trees is show in Figure 2. The task is to develop a classification rule that can determine the class of any object from its values of the attributes[6]. Random Forests are an ensemble method which builds several decision trees and merges them together in order to achieve a better accuracy on the prediction results.

OBJECTIVES AND METHODOLOGY

Two 49-points Datasets from two different 7x7 mortar grids were obtained from Nanoindentation experiments. The nanomechanical assessment was conducted in a Hysitron TriboLab® Nanomechanical Test Instrument, which allows the application of loads from 1 μN to 30 mN and records the displacement as a function of applied loads with a high load resolution (1 nN) and a high displacement resolution (0.04 nm). In all measurements, a total of 10 indents are averaged so as to determine the mean hardness (H) and elastic modulus (E) values for statistical reasons, with a 50 μm spacing, in a clean area environment with 45 % humidity and 23 °C ambient temperature. In order to operate under closed loop control, feedback control option was selected. All measurements have been performed at 200nm of displacement (so as for the indentation response not to be affected by discrete nature of colloidal particles and microstructure/interaction of different phases, using the standard three-sided pyramidal Berkovich tip indenter, with an average curvature radius of 100 nm[10,11].

Regarding the first Dataset, the mortar grid was also studied under the Optical Microscope in order to map each measurement point to its surface colour. Via this, each Nanoindentation measurement will be mapped to a specific colour. Later, this Dataset will be used as training set for a Random Forest classifier in order to learn how the colour (label) behaves depending on E and H values. The trained model will be used for predicting the surface colour of each measurement point on the second 7x7 Dataset.

In addition, K-Means algorithm will be applied on the second 7x7 Dataset and cluster the Data of E and H values. Via this, the constituent phase of each point in the grid will be estimated and cross-validated with the bibliography values. Furthermore, as mentioned above, the Random Forest classifier from previous step will be used for predicting the colour of each point on the second grid. Finally, possible correlations between colour (optical property) and the resulted constituent phase of K-Means clustering will be studied.

RESULTS AND DISCUSSION

K-Means was applied on the second 49-measurements grid in order to extract information about the underlying structure of the data. The optimum amount of clusters was set as 7. In Figure 3, resulted Clusters from K-Means on the second grid are presented; each cluster signifies a constituent phase of the mortar.

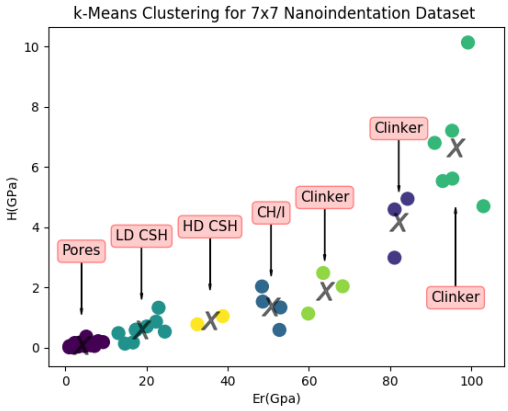

Figure 3. Resulted Clusters from K-Means on 2nd mapping grid. Each cluster signifies a constituent phase of the mortar

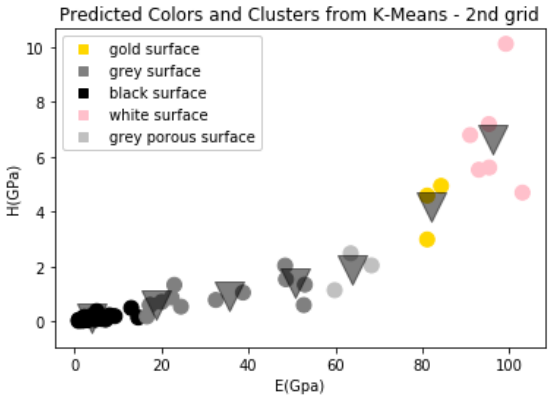

Figure 4. Predicted surface colours for each measurement of the 2nd mapping grid from Random Forests algorithm

In this case, the resulted clusters are 7 and cover a bigger range of E values. As a result, clinker phases are also expected. Again, the first cluster takes values between 0.2 and 15 GPa and corresponds to the pores inside the mortar. LD C-S-H includes as expected values between 16 and 25 GPa. The third cluster (containing 2 points) shows the HD C-S-H phase and its E values are in the range between 30 and 40 GPa. The fourth cluster refers to CH/I and has as expected points with E values between 45 and 58 GPa. Clinker phases are the last 3 clusters and take values between 60 and 105 GPa. Clinker can be clustered into 3 groups: one with E values between 60 and 75 GPa, one with E values between 80 and 90 GPa and one with E values above 90 GPa. It is clear, that K-Means shows similarities in the cluster ranges for both grids although the number of total clusters is not the same. Furthermore, via this, the various ranges of E values have been assigned to a specific phase according to the bibliography and this has been validated throughout K-Means. In table 1 below, literature and this work's values based on clustering method are presented.

Table 1. E and H values from literature[12,13-16] and K-Means clustering for each mortar phase

Phases	**E (GPa)**		**H (GPa)**	
	Literature	*Cluster (this work)*	*Literature*	*Cluster (this work)*
Pores	0 – 13	0.2 – 15	0.16 – 0.18	0.1– 0.35
LD C-S-H	13 – 26	16 – 26	0.4 – 0.8	0.4 GPa – 1.8
HD C-S-H	26 -39	26 -40	0.8 – 1.25	0.4 - 2.1

CH-CH/I	35.1 – 42.9	41 – 58	1.31 – 1.66	0.8 - 3.2
Clinker	-	>60	-	>1.8

In Figure 4, the predicted colours from random forests algorithm in the 2nd grid are presented. The 7 points (triangles) signify the clusters from K-Means algorithm of Figure 3. Random forests algorithm predicted measurements with E values lower than 18 GPa as black surface. Points between 18 GPa and 58 GPa were predicted as grey surface. Via this, the 1st cluster of Figure 3 (pores) can be correlated with the black surface. In addition, the clusters 2,3,4 from Figure 3 (LD CSH, HD CSH, CH/I) can be correlated with the gray coloured surface. Points with values between 60 GPa and 71 GPa were predicted as grey porous surface, which can be correlated with the cluster "Clinker" of Figure 3. Likewise, gold surface (80-88 GPa) and white surface (>90GPa) can also be correlated with cluster "Clinker" of Figure 3. Therefore, the cluster "Clinker" of Figure 3 contains points with 3 possible colours (grey porous, gold, white). In Table 2, resulted ranges for E, H and plasticity (integrated area of load-unload curve) values for each surface colour are presented.

Table 2. Resulted ranges for E, H values and plastic deformation for each surface colour

Surface Colour	E (GPa)	H (GPa)
black	<18	< 0.4
grey	18 - 58	0.5 – 3.1
grey porous	60 - 71	1 – 2.2
gold	80 - 88	2.4 - 5
White	> 90	> 4.2

In Figure 5 the reconstruction of the constituent phases of resulted K-Means (Figure 3) in the 2nd grid is presented. Deep blue colours signify pores in the corresponding areas. Light blue corresponds to LD CSH areas, while HD CSH areas are the deeper green areas and CH/I areas correspond to the light green colours (yellow areas correspond to Clinker).

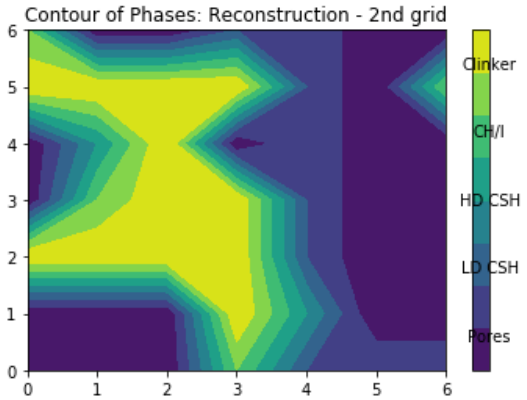

Figure 5. Reconstruction of constituent phases for the 2nd mapping grid dataset from elastic modulus

CONCLUSIONS

In this work, constituents phase reconstruction through applied machine learning in nanoindentation mapping data of mortar surface was conducted. Machine Learning algorithms were used to cluster the data and evaluate the phase of each measurement; additionally, possible correlation between optical properties (color) and the clustered phase from the first step is assessed. For the second part, training a model from other labeled Data was required. For this purpose, another Nanoindentation Dataset with labeled with surface color for every instance used as training set in order to construct the classifier (using Random Forests algorithm). This classifier was then used to predict the surface color for each measurement of the two 7x7 grids. Finally, contours were created in order to reconstruct the constituent phase of the mortar grid just from E and H measurements.

The resulted clusters from K-Means were cross-validated with bibliography range values of constituent phases and as proved, each K-Means cluster approximates a single constituent phase of mortar. Furthermore, the constituent phases of the mortar grid are correlated with the surface color, as each K-Means cluster corresponds to a unique surface color as predicted from Random Forests.

BIBLIOGRAPHY

[1] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K. M. Ho, I. Takeuchi, On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. Scientific reports, 4 (2014), 6367.
[2] T. Mueller, A. G. Kusne, R. Ramprasad, Machine learning in materials science: Recent progress and emerging applications. Reviews in Computational Chemistry, 29 (2016), 186-273.
[3] S. Chang, T. Cohen, B. Ostdiek, (2018). What is the machine learning?. Physical Review D, 97(5), 056009.
[4] J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, J. (2018). Constraining Effective Field Theories with Machine Learning. arXiv preprint arXiv:1805.00013.
[5] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects. npj Computational Materials, 3(1) (2017), 54.
[6] J. Nieves, I. Santos, Y. K. Penya, S. Rojas, M. Salazar, P. G. Bringas, Mechanical properties prediction in high-precision foundry production. In Industrial Informatics, 2009. INDIN 2009. 7th IEEE International Conference on (pp. 31-36). IEEE (2009, June)., https://doi.org/10.1109/INDIN.2009.5195774
[7] A. K. Jain, Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8) (2010), 651-666., https://doi.org/10.1016/j.patrec.2009.09.011
[8] M. Awad, R. Khanna, Efficient learning machines: theories, concepts, and applications for engineers and system designers, Apress (2015)
[9] J. R. Quinlan, Induction of decision trees. Machine learning, 1(1) (1986), 81-106.
[10] E.P. Koumoulos, D.A. Dragatogiannis, C.A. Charitidis, Nanomechanical properties and deformation mechanism in metals, oxides and alloys. In Nanomechanical Analysis of High-Performance Materials (pp. 123-152). Springer (2014), Dordrecht.
[11] E. P. Koumoulos, C. A. Charitidis, D. P. Papageorgiou, A.G. Papathanasiou, A. G. Boudouvis, Nanomechanical and nanotribological properties of hydrophobic fluorocarbon dielectric coating on tetraethoxysilane for electrowetting applications. Surface and Coatings Technology, 206(19-20) (2012), 3823-3831., https://doi.org/10.1016/j.surfcoat.2012.01.034
[12] Y. Gao, C. Hu, Y. Zhang, Z. Li, J. Pan, Characterisation of the interfacial transition zone in mortars by nanoindentation and scanning electron microscope. Magazine of Concrete Research (2018), 1-8., https://doi.org/10.1680/jmacr.17.00161
[13] G. Constantinides, F. J. Ulm, The nanogranular nature of C–S–H. Journal of the Mechanics and Physics of Solids, 55(1) (2007), 64-90., https://doi.org/10.1016/j.jmps.2006.06.003
[14] W. Zhu, J.J. Hughes, JN. Bicanic, C. J. & Pearce, Nanoindentation mapping of mechanical properties of cement paste and natural rocks. Materials characterisation, 58(11-12) (2007), 1189-1198., https://doi.org/10.1016/j.matchar.2007.05.018
[15] C. Hu, Nanoindentation as a tool to measure and map mechanical properties of hardened cement pastes. MRS Communications, 5(1) (2015), 83-87., https://doi.org/10.1557/jmr.2009.0149
[16] T. Howind, T., J. J. Hughes, W. Zhu, F. Puertas, S. Goñi Elizalde, M. S, Hernandez, J. S. Dolado, Mapping of mechanical properties of cement paste microstructures., 2011