Supplementary material for: On the nature of mixed-type features in materials datasets

Duy-Tai Dinh[1], Duong-Nguyen Nguyen[1], Hieu-Chi Dam[1,2]
[1] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[2] *National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*

(Dated: 28 January 2021)

I. AN EXAMPLE

Table I presents a mixed-type material dataset of 20 binary compounds, as described by 12 variables. The categorical variables comprise the atomic numbers of A and B (Z_A, Z_B) and types of elements (type$_A$, type$_B$), whereas the numerical variables consist of continuous variables such as unit-cell volume (V_{cell}), electronegativity of A and B (n_{eA} and n_{eB}), ionization potential (IP_A), boiling temperature (T_{bA}), melting temperature (T_{mA}) of the element A, atomic radius of the element B (r_B). The formation energy (E_{form}) is considered as the physical property of interest.

This example illustrates how the model works on the dataset shown in Table I. The model uses categorical and numerical variables as their nature for the task of clustering and regression, respectively. First, it uses K-cat to partition the instances in categorical part of the data into two clusters (Fig. 1a). Using the labels of instances in each cluster, the model splits the instances in numerical part of the data into two groups D_1 and D_2, which also contain the observed values of E_{form}. In the next step, it builds the lasso regression to estimate predicted values for instances in each group. The R^2 and MAE are then determined by all observed and predicted values of E_{form} for the whole dataset. Fig. 1b shows the confusion matrix representing the correlation between observed and predicted target values. For all sub-figures, the horizontal and vertical axes depict the observed and predicted target values, respectively. The upper left subfigure shows the correlation between observed values of instances in D_1 and predicted values produced by the model M_1, using both predicting and observed variables in D_1. The upper right subfigure shows the correlation between observed values of instances in D_2 and predicted values obtained by model M_1. It can be seen that the observed and predicted values are highly different. Similarly, the bottom right sub-figure shows the correlation between observed values of instances in D_2 and predicted values obtained by model M_2, fitted by using both predicting and observed variables in D_2. The bottom left sub-figure shows the correlation between observed values of instances in D_1 and predicted values obtained by model M_2. Generally, for all instances in the dataset, the model achieves 0.865 and 0.146 (Å) for R^2 and MAE, respectively.

II. FEATURE USAGE

A. The Lattice dataset of body-centred cubic structure data

The Lattice dataset contains 1,439 binary AB body-centred cubic crystals[1,2]. The lattice constant is considered as the property of interest. The features are classified into two parts:

- The numerical variables consist of the quantitative properties of the binary AB elements: unit-cell density (p), the difference in electronegativity (d_χ), covalent radius of A and B (r_{covA}, r_{covB}), mass (m_A, m_B), the density of atoms per unit volume (p_A, p_B), the sum of the atomic orbital B and the difference in the electronegativity (Sum$_{AD}$).
- The categorical variables consist of the quantitative properties of the binary AB: the atomic numbers of A and B (Z_A, Z_B), the atomic orbital of A and B (l_A, l_B), atomic names of A and B (atom$_A$, atom$_B$), groups (group$_A$, group$_B$), periods of A and B (period$_A$, period$_B$), group index.

Looking in more detail at this dataset, the atomic numbers of A are in the set of {3, 11-14, 22-24, 26-31, 33, 44-47, 74, 76-79}, whereas the atomic numbers of B are in the set of {1-42, 44-57, 72-83}. The atomic orbital of A or B receives each value in {0, 1, 2, 3}. A belongs to a group in the set {1, 2, 4-6, 8-15}, while B falls into a group in the range of 1-18. In addition, A atoms have periods in the range of 2-6, while B atoms have periods in the range of 1-6.

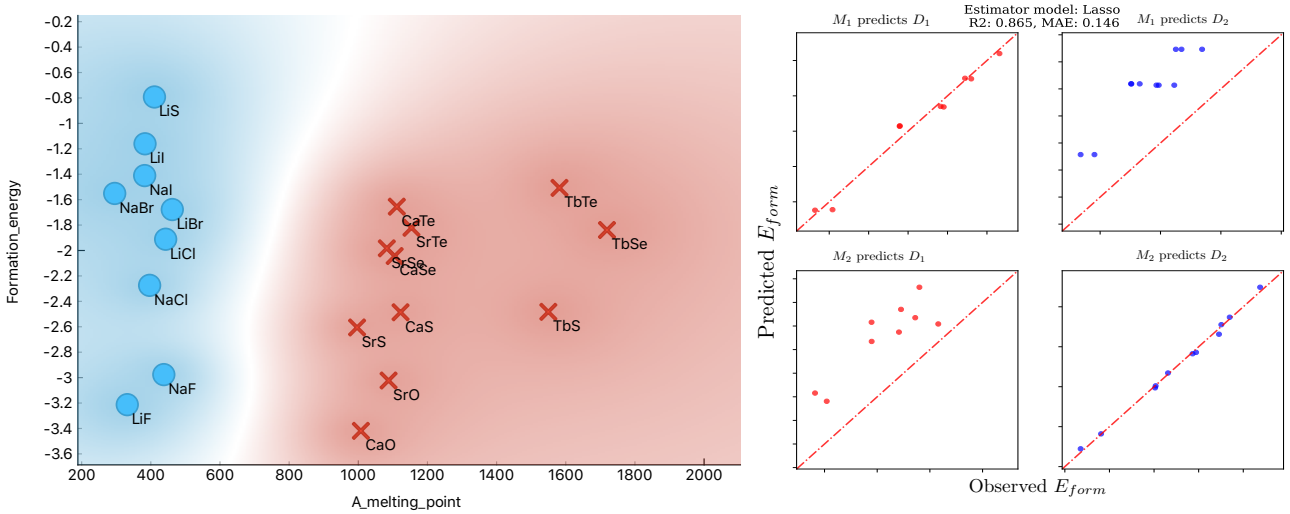
B. The TC dataset of rare earth-transition metal alloys

The TC dataset was collected from NIMS Atomwork database[2,3]. It contains 101 binary alloys of transition and rare earth metals. The Curie temperature (T_C) is considered as the property of interest. The features are split into two parts:

- The numerical variables consist of the atomic properties of the transaction metal elements (T) and rare earth elements (R): atom radius of T and R (r_T, r_R), covalent radius (r_{covT}, r_{covR}), ionization potential (IP_T, IP_R), electronegativity (χ_T,

TABLE I: A running example chemical dataset

Id \ A_j	Z_A	Z_B	type$_A$	type$_B$	V_{cell}	n_{eA}	n_{eB}	IP_A	T_{bA}	T_{mA}	r_B	E_{form}
NaF	11	9	Alkali	Halogen	25.894	0.93	3.98	5.139	1156	370.87	0.5	-2.958
NaCl	11	17	Alkali	Halogen	46.096	0.93	3.16	5.139	1156	370.87	1	-2.105
NaBr	11	35	Alkali	Halogen	54.749	0.93	2.96	5.139	1156	370.87	1.15	-1.587
NaI	11	53	Alkali	Halogen	69.675	0.93	2.66	5.139	1156	370.87	1.4	-1.278
SrO	38	8	Alkaline	Nonmetal	35.277	0.95	3.44	5.695	1655	1050	0.6	-3.094
SrS	38	16	Alkaline	Nonmetal	55.732	0.95	2.58	5.695	1655	1050	1	-2.482
SrSe	38	34	Alkaline	Nonmetal	62.626	0.95	2.55	5.695	1655	1050	1.15	-2.068
SrTe	38	52	Alkaline	Nonmetal	76.007	0.95	2.1	5.695	1655	1050	1.4	-1.746
LiF	3	9	Alkali	Halogen	17.022	0.98	3.98	5.392	1615	453.69	0.5	-3.18
LiS	3	16	Alkali	Nonmetal	31.689	0.98	2.58	5.392	1615	453.69	1	-0.841
LiCl	3	17	Alkali	Halogen	34.202	0.98	3.16	5.392	1615	453.69	1	-2.107
LiBr	3	35	Alkali	Halogen	41.899	0.98	2.96	5.392	1615	453.69	1.15	-1.548
LiI	3	53	Alkali	Halogen	54.697	0.98	2.66	5.392	1615	453.69	1.4	-1.199
CaO	20	8	Alkaline	Nonmetal	28.332	1	3.44	6.113	1757	1115	0.6	-3.323
CaS	20	16	Alkaline	Nonmetal	46.695	1	2.58	6.113	1757	1115	1	-2.488
CaSe	20	34	Alkaline	Nonmetal	53.051	1	2.55	6.113	1757	1115	1.15	-2.029
CaTe	20	52	Alkaline	Nonmetal	65.451	1	2.1	6.113	1757	1115	1.4	-1.652
TbS	65	16	Lanthanoid	Nonmetal	42.813	1.1	2.58	5.86	3503	1629	1	-2.344
TbSe	65	34	Lanthanoid	Nonmetal	48.63	1.1	2.55	5.86	3503	1629	1.15	-1.772
TbTe	65	52	Lanthanoid	Nonmetal	59.125	1.1	2.1	5.86	3503	1629	1.4	-1.311



(a) Clustering results on the running example dataset

(b) Regression models fitting on two different clusters

FIG. 1: Clustering and regression based prediction model on the running example dataset. (1a) shows clustering results produced by the K-cat algorithm ; (1b) shows the confusion matrix representing the correlation between observed and predicted target values when two models predicts their own groups (diagonal) and different groups (off-diagonal).

χ_R), boiling temperatures (T_{bT}, T_{bR}), total spin quantum numbers (S_{3d}, S_{4f}), total orbital angular momentum quantum numbers (L_{3d}, L_{4f}), total orbital angular momentum (J_{3d}, J_{4f}), the strong spin-orbit coupling effect for R metallic elements $J_{4f}g_j$ and $J_{4f}(1-g_j)$, the distance between two transition metal elements (d_{TT}), between two rare earth elements (d_{RR}), between transition metal and rare earth elements (d_{TR}), the mean radius of the unit cell between two transition metal elements ($mean_{rTT}$), between two rare earth elements ($mean_{nRR}$), between transition metal and rare earth elements ($mean_{nTR}$), the concentrations of T and R (C_T and C_R) in units of atoms $\mathring{A}^{-3}$.

- The categorical variables describe the remaining atomic properties of T and R elements. They are T and R elements in each compound (R_mental, T_mental), the atomic numbers of T and R (Z_T, Z_R).

C. The Octet Binary dataset

The octet binary materials dataset has 82 materials obtained with DFT using the local-density approximation (LDA) for the exchange-correlation interaction[4]. It contains crystal structure of binary compound semiconductors, which are known to crystallize in zinc blende (ZB), wurtzite (WZ), or rocksalt (RS) structures. The difference in LDA energy between RS and ZB ($\Delta E = E(RS) - E(ZB)$) in eV is considered as the physical property of interest. The features are split into two parts as follows:

- The numerical variables describe quantitative properties of two elements: the lattice constant a_0 for the three considered crystal structure ($a_0(RS)[\mathring{A}]$, $a_0(ZB)[\mathring{A}]$, $a_0(WZ)[\mathring{A}]$), the predicted ΔE for the $1D$, $2D$, $3D$ descriptors ($\Delta E_{1D}[eV]$, $\Delta E_{2D}[eV]$, $\Delta E_{3D}[eV]$), the value of the $1D$, $2D$ and $3D$ descriptors ($d_1[eV\mathring{A}^{-2}]$, $d_2[\mathring{A}]$, $d_3[\mathring{A}]$), and the difference in energy between ZB and WZ structures ($\Delta E(WZ) = E(ZB) - E(WZ)$).

- The categorical variables consist of the physical property atomic numbers of A and B elements (Z_A,Z_B).

Looking in more detail at the dataset, the atomic numbers of A are in the ranges of 3-6, 11-14, 19-20, 29-32, 37-38, 47-50, 55-56, whereas the atomic numbers of B are in the ranges of 6-9, 14-17, 32-35, 50-53.

D. The Fm$\bar{3}$m binary compounds dataset

The AB materials dataset contains 239 binary compounds of the form $Fm\bar{3}m$ collected from the Materials Project[5]. The formation energy as the physical property of interest. The features are split into two parts:

- The numerical variables consist of the quantitative properties of both A and B atoms: volume, band gap, atomic radius (r_A, r_B), average ionic radius (r_{ionA}, r_{ionB}), ionization potential (IP_A, IP_B), electronegativity (χ_A, χ_B), boiling temperatures (T_{bA}, T_{bB}), melting temperatures (T_{mA}, T_{mB}), valance electrons (V_{eA}, V_{eB}), first ionization energy (I_{1A}, I_{1B}).

- The categorical variables consist of the nonnumerical values and nominal values in the form of qualitative properties of both A and B atoms: atomic numbers (Z_A,Z_B), types (type$_A$, type$_B$), elements (element$_A$, element$_B$), rows (row$_A$, row$_B$), groups (group$_A$, group$_B$), blocks (block$_A$, block$_B$).

Looking in more detail at this dataset, the domains of Z_A and Z_B are finite, contain 60 and 13 values over 239 compounds for atoms A and B, respectively. Regarding element types (type$_A$, type$_B$), the A atoms are classified into nine groups representing most of the metallic forms such as alkali, alkaline, lanthanides, metalloid, post-transition and transition metals, whereas the B atoms are classified into five groups representing for halogen, non-metals and metalloids. Moreover, the compounds are formed by different combinations of A and B elements, corresponding with Z_A and Z_B. The row$_A$ and row$_B$ features provide the information of rows where A and B are in the periodic table. Specifically, all A elements spread on six rows (2, 3, 4, 5, 6, 8) and B elements spread on four rows (2, 3, 4, 5) of the periodic table. Finally, block$_A$ and block$_B$ provide the block information of A and B in the periodic table, where A comes from blocks s, p, d, f, and B comes from block p.

[1] K. Takahashi, L. Takahashi, J. D. Baran, and Y. Tanaka, "Descriptors for predicting the lattice constant of body centered cubic crystal," The Journal of chemical physics **146**, 204104 (2017).
[2] D.-N. Nguyen, T.-L. Pham, V.-C. Nguyen, T.-D. Ho, T. Tran, K. Takahashi, and H.-C. Dam, "Committee machine that votes for similarity between materials," IUCrJ **5**, 830–840 (2018).
[3] Y. Xu, M. Yamazaki, and P. Villars, "Inorganic materials database for exploring the nature of material," Japanese Journal of Applied Physics **50**, 11RH02 (2011).
[4] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: critical role of the descriptor," Physical review letters **114**, 105503 (2015).
[5] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," Apl Materials **1**, 011002 (2013).