On the nature of mixed-type features in materials datasets : Supplemetary Information

Duy-Tai Dinh, Duong-Nguyen Nguyen, Hieu-Chi Dam*

[a]*International Excellent Core for Materials Informatics, Headquarters for Excellent Core Promotion, Japan Advanced Institute of Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

1. Distinguishing between numerical and categorical features

The AB materials dataset

Table 1: List of numerical and categorical feature in the AB materials dataset

Numerical features	Categorical features
volumne, Band_gap, A_e_negativity, A_valence_e, A_first_ionization, A_boiling_point, A_melting_point, A_atomic_radius, A_average_ionic_radius, B_e_negativity, B_valence_e, B_first_ionization, B_boiling_point, B_melting_point, B_atomic_radius, B_average_ionic_radius, **Formation_energy**	A_Z, B_Z, A_type, B_type, A_element, B_element, A_row, A_group, A_block, B_row, B_group, B_block

The Octet Binary dataset

Table 2: List of numerical and categorical feature in the Binary octet dataset

Numerical features	Categorical features
a0(RS), a0(ZB), delta-E1D, delta-E2D, delta-E3D, d1, d2, d3, a0, delta-EWZ, **delta-E**	ZA, ZB

*Corresponding Author

Email addresses: taidinh@jaist.ac.jp (Duy-Tai Dinh), nguyen@jaist.ac.jp (Duong-Nguyen Nguyen), dam@jaist.ac.jp (Hieu-Chi Dam)

The Infrared Intensity (IR) dataset

Table 3: List of numerical and categorical feature in the IR dataset

Numerical features	Categorical features
gap, maxp, minp, p_a, p_alpha, c_a, c_alpha, Relaxed energy, p_b, p_beta, c_b, c_beta, SCF bandgap (eV), p_c, p_gamma, c_c, c_gamma, Density , Volume, hsk based bandgap, dk based bandgap, Clarke, Cahill, **Formation energy**	formula, spg, Crystal system, Point group, Dimensionality, nAtoms_conv, nAtoms_prim

The Lattice dataset of body-centred cubic structure data

Table 4: List of numerical and categorical feature in the Lattice dataset

Numerical features	Categorical features
Density, Atomic_Orbital_A, Atomic_Orbital_B, diff_elecneg_A_B, atom_orbit_B_plus_diff_elecneg_A_B, cov_r_A, cov_r_B, mass_A, mass_B, density_A, density_B, **Lattice_Constant**	Atomic_Number_A, Atomic_Number_B, atom_A, atom_B, group_A, group_B, period_A, period_B, group_index

The Phonon dataset

Table 5: List of numerical and categorical feature in the Phonon dataset

Numerical features	Categorical features
a_(A), b_(A), c_(A), density(g_cm^3), volume(A^3), energy_atom(eV), volume_atom(A^3), ave_mass, min_max_mass, ave_chem, ave_elec, band_gap(eV), **energy(eV)**	nelements, nsites, nbranch, spacegroup, max_atom_no, min_atom_no, max_row, min_row, max_col, min_col, max_mass, min_mass, max_chem, min_chem, max_elec, min_elec, alpha, beta, gama, ave_atom_no, max_radii, min_radii, ave_radii, min_max_radii, ave_row, ave_col

The Superconductivity dataset

Table 6: List of numerical and categorical feature in the Superconductivity dataset

Numerical features	Categorical features
wtd_mean_atomic_mass, wtd_gmean_atomic_mass, wtd_entropy_atomic_mass, wtd_range_atomic_mass, wtd_std_atomic_mass, wtd_mean_fie, wtd_gmean_fie, wtd_entropy_fie, wtd_range_fie, wtd_std_fie, wtd_mean_atomic_radius, wtd_gmean_atomic_radius, wtd_entropy_atomic_radius, wtd_range_atomic_radius, wtd_std_atomic_radius, wtd_mean_Density, wtd_gmean_Density, wtd_entropy_Density, wtd_range_Density, wtd_std_Density, wtd_mean_ElectronAffinity, wtd_gmean_ElectronAffinity, wtd_entropy_ElectronAffinity, wtd_range_ElectronAffinity, wtd_std_ElectronAffinity, wtd_mean_FusionHeat, wtd_gmean_FusionHeat, wtd_entropy_FusionHeat, wtd_range_FusionHeat, wtd_std_FusionHeat, wtd_mean_ThermalConductivity, wtd_gmean_ThermalConductivity, wtd_entropy_ThermalConductivity, wtd_range_ThermalConductivity, wtd_std_ThermalConductivity, wtd_mean_Valence, wtd_gmean_Valence, wtd_entropy_Valence, wtd_range_Valence, wtd_std_Valence, **critical_temp**	number_of_elements, mean_atomic_mass, gmean_atomic_mass, entropy_atomic_mass, range_atomic_mass, std_atomic_mass, mean_fie, gmean_fie, entropy_fie, range_fie, std_fie, mean_atomic_radius, gmean_atomic_radius, entropy_atomic_radius, range_atomic_radius, std_atomic_radius, mean_Density, gmean_Density, entropy_Density, range_Density, std_Density, mean_ElectronAffinity, gmean_ElectronAffinity, entropy_ElectronAffinity, range_ElectronAffinity, std_ElectronAffinity,mean_FusionHeat, gmean_FusionHeat, entropy_FusionHeat, range_FusionHeat, std_FusionHeat, mean_ThermalConductivity, gmean_ThermalConductivity, entropy_ThermalConductivity, range_ThermalConductivity, std_ThermalConductivity, mean_Valence, gmean_Valence, entropy_Valence, range_Valence, std_Valence

The TC dataset of rare earth-transition metal alloys

Table 7: List of numerical and categorical feature in the TC dataset

Numerical features	Categorical features
Z_R, Z_T, Cov_r_T, Elec_neg_T, L_3d, J_3d, RR_distance, RT_distance, TT_distance, mean_nTR, mean_nRR, mean_nRT, C_T, C_R, **Tc**	R_metal, T_metal, r_R, cov_r_R, Boiling_point_R, Elect_neg_R, Ion_poten_R, S_4f, L_4f, J_4f, gj_4f, J_4f_x_gj_4f, J_4f_x(gj_4f-1), r_T, Boiling_point_T, Ion_poten_T, S_3d

The Thermoelectric dataset

Table 8: List of numerical and categorical feature in the Thermoelectric dataset

Numerical features	Categorical features
volumne, Band_gap, A_e_negativity, A_valence_e, A_first_ionization, A_boiling_point, A_melting_point, A_atomic_radius, A_average_ionic_radius, B_e_negativity, B_valence_e, B_first_ionization, B_boiling_point, B_melting_point, B_atomic_radius, B_average_ionic_radius, **ZT**	Atom_number, space_group, max_Z,min_Z, min_eleg, max_mass, min_mass, max_covrad, max_period, mean_period, min_group, Host, Dopant, X_element, Z_H, Z_D, Z_X, Z_S, Elec_neg_H, Elec_neg_D, Elec_neg_X, Elec_neg_S, Ion_potent_H, Ion_potent_D, Ion_potent_X, Ion_potent_S, Radius_H, Radius_D, Radius_X, Radius_S

The 2D-Thermoelectric dataset

Table 9: List of numerical and categorical feature in the 2D-Thermoelectric dataset

Numerical features	Categorical features
OPT_SCF_gap, n-Seebeck, p-Seebeck, n-powerfact, p-powerfact, n-cond, p-cond, p-ZT, **n-ZT**	Spg_numb, Spg_symb, Cryst, Dimensionality

The Solar cell materials dataset

Table 10: List of numerical and categorical feature in the Solar cell materials dataset

Numerical features	Categorical features
mean_Number, dev_Number, mean_MendeleevNumber, dev_MendeleevNumber, mean_AtomicWeight, maxdiff_AtomicWeight, dev_AtomicWeight, max_AtomicWeight, min_AtomicWeight, most_AtomicWeight, mean_MeltingT, maxdiff_MeltingT, dev_MeltingT, mean_CovalentRadius, dev_CovalentRadius, mean_Electronegativity, dev_Electronegativity, mean_GSvolume_pa, dev_GSvolume_pa, dev_SpaceGroupNumber, frac_sValence, frac_pValence, frac_dValence, frac_fValence, MeanIonicChar, **current_known_FE**	NComp, Comp_L2Norm, Comp_L3Norm, Comp_L5Norm, Comp_L7Norm, Comp_L10Norm, maxdiff_Number, max_Number, min_Number, most_Number, maxdiff_MendeleevNumber, max_MendeleevNumber, min_MendeleevNumber, most_MendeleevNumber, max_MeltingT, min_MeltingT, most_MeltingT, mean_Column, maxdiff_Column, dev_Column, max_Column, min_Column, most_Column, mean_Row, maxdiff_Row, dev_Row, max_Row, min_Row, most_Row, maxdiff_CovalentRadius, max_CovalentRadius, min_CovalentRadius, most_CovalentRadius, maxdiff_Electronegativity, max_Electronegativity, min_Electronegativity, most_Electronegativity, mean_NsValence, maxdiff_NsValence, dev_NsValence, max_NsValence, min_NsValence, most_NsValence, mean_NpValence, maxdiff_NpValence, dev_NpValence, max_NpValence, min_NpValence, most_NpValence, mean_NdValence, maxdiff_NdValence, dev_NdValence, max_NdValence, min_NdValence, most_NdValence, mean_NfValence, maxdiff_NfValence, dev_NfValence, max_NfValence, min_NfValence, most_NfValence, mean_NValance, maxdiff_NValance, dev_NValance, max_NValance, min_NValance, most_NValance, mean_NsUnfilled, maxdiff_NsUnfilled, dev_NsUnfilled, max_NsUnfilled, min_NsUnfilled, most_NsUnfilled, mean_NpUnfilled, maxdiff_NpUnfilled, dev_NpUnfilled, max_NpUnfilled, min_NpUnfilled, most_NpUnfilled, mean_NdUnfilled, maxdiff_NdUnfilled, dev_NdUnfilled, max_NdUnfilled, min_NdUnfilled, most_NdUnfilled, mean_NfUnfilled, maxdiff_NfUnfilled, dev_NfUnfilled, max_NfUnfilled, min_NfUnfilled, most_NfUnfilled, mean_NUnfilled, maxdiff_NUnfilled, dev_NUnfilled, max_NUnfilled, min_NUnfilled, most_NUnfilled, maxdiff_GSvolume_pa, max_GSvolume_pa, min_GSvolume_pa, most_GSvolume_pa, mean_GSbandgap, maxdiff_GSbandgap dev_GSbandgap, max_GSbandgap, min_GSbandgap, most_GSbandgap, mean_GSmagmom, maxdiff_GSmagmom, dev_GSmagmom, max_GSmagmom, min_GSmagmom, most_GSmagmom, mean_SpaceGroupNumber, maxdiff_SpaceGroupNumber, max_SpaceGroupNumber min_SpaceGroupNumber, most_SpaceGroupNumber, CanFormIonic, MaxIonicChar