

Supplementary Information

Small Data Machine Learning in Materials Science

Pengcheng Xu^a, Xiaobo Ji^b, Minjie Li^{b,*}, Wencong Lu^{a,b,*}

^a Materials Genome Institute, Shanghai University, Shanghai 200444, China

^b Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China

AUTHOR INFORMATION

Corresponding Author E-mail: Minjie Li*: minjieli@shu.edu.cn, Wencong Lu*: wclu@shu.edu.cn

ORCID Wencong Lu: 0000-0001-5361-6122

Supplementary table 1: Commonly used software and platform of text mining

Software	Description	ULR
ActivePoint	ActivePoint could provide natural language processing and intelligent online catalogs	http://www.activepoint.com/
Basis Technology	Natural language processing techniques for analyzing unstructured multilingual texts	https://www.basistech.com/
Copernic Summarizer	Ability to read and summarize text content of documents and web pages from different applications	https://copernic.com/en/
DiscoverText	Cloud-based text analytics solution with many powerful features, including an active learning classification engine	https://discovertext.com/
IBM SPSS Predictive Analytics	Suite for data and text mining	https://www.ibm.com/spss
ISYS Search Software	A search software focused on embedded search, text extraction, federated access solutions and text analysis	https://www.hyland.com/en/perceptive
KNIME	An open source analysis platform that provides extensions to current text analysis software, including Stanford	https://www.knime.com/

NLP, Palladin, and Linguamatics

Lexalytics	Lexalytics provides enterprise and hosted text analytics software that converts unstructured text into structured data	https://www.lexalytics.com/
Megaputer Text Analyst	Megaputer Text Analyst provides semantic analysis, summarization, clustering, navigation and natural language retrieval of free-form text	https://www.megaputer.com/
MonkeyLearn	Text mining tools for creating machine learning applications, capable of providing classification, extraction, clustering and regression modules via web and API	https://monkeylearn.com/
Ontotext	Hybrid text mining, inference and graphical databases with semantic technologies to provide optimized knowledge management, search and semantic analysis solutions	https://www.ontotext.com/
VP Student Edition	Powerful text mining and visualization tools for discovering knowledge from scientific literature and other field-structured text databases	https://vpinstitute.org/vp-marketplace/
Aika	An open source library for frequent pattern mining of text, using the ideas of neural networks and syntactic induction	http://aika.network/
Coding Analysis Toolkit (CAT)	Free, open source, web-based text analysis tool	http://cat.texifter.com/

Supplementary table 2: Materials databases and corresponding introduction

Database	Description	ULR
Materials Project	The Materials Project provides structural information and properties of more than 130,000 inorganic compounds, including first-principles calculations of band structures, elastic tensors, piezoelectric tensors, and other	https://www.materialsproject.org/

	properties.	
OQMD	OQMD is a computational thermodynamic and structural database based on density functional theory. The materials in the database include battery materials, hydrogen storage materials, solar materials and thermoelectric materials. Additionally, OQMD offers a machine learning toolkit.	http://oqmd.org/
NIST	NIST database is a comprehensive mass spectrometry database that can be used for metabolite identification based on LC-MS metabolomics, covering almost all material systems	https://www.nist.gov/srd/
ICSD	ICSD has been published since 1913 and has included nearly 100,000 compounds. The database includes chemical formulas, unit cell parameters, space groups, atomic coordinate thermal parameters and other information. All data are recorded by experts and revised several times	http://icsd.fiz-karlsruhe.de/
MPDS	MPDS is the most comprehensive inorganic material database in the world. The database includes material literature published around the world from 1900 to the present, and has collected more than one million experimental and computational data. The material properties covered by MPDS include mechanics, thermodynamics, electromagnetism, optics, etc.	https://asm.mpds.io
NOMAD	NOMAD has established a comprehensive database dedicated to collecting all computational material data, and has now	https://nomad-coe.eu

	included millions of computational results for open use	
AFLOWlib	AFLOWlib stores more than 3.56 million material structures including inorganic compounds, binary alloys and multi-component alloys, and 700 million first-principles calculated material properties data, which is the largest data volume among many databases.	http://aflowlib.org/
Materials Genome Engineering Databases	The largest material genetic engineering database platform of China. In addition to the database, the platform also has a first-principles online calculation engine, atomic potential function library, online data mining system and many other functions	https://www.mgedata.cn/
OMDB	OMDB is a three-dimensional organic crystal electronic structure database containing the electronic structures, density of states and other properties of pure organic and organometallic compounds known to date.	http://omdb.diracmaterials.org
CSD	A database of crystal structures of small-molecule organics and metal-organic compounds created by CSD in 1965. The database includes crystal structure data of 1.15 million small-molecule organic compounds and metal-organic compounds from the literature, including unit cell parameters, atomic coordinates, and citations.	https://www.ccdc.cam.ac.uk/
QM9	QM9 provides quantum chemical properties for a relevant, consistent, and comprehensive chemical space of small organic molecules.	https://paperswithcode.com/dataset/qm9
High entropy alloy	TCHEA6 is the thermodynamic and properties	https://thermocalc.com/products/d

database	database designed for use with High Entropy Alloys (HEAs), also known as Multi-principal Element Alloys (MPEA).	atabases/high-entropy-alloys/
Starrydata2	Starrydata2 can automatically extract numerical data from the mapping of scientific papers and the chemical composition of the corresponding samples, and has successfully collected experimental data from over 11,500 thermoelectric material samples from the mapping images.	https://www.starrydata2.org/
Novamag database	Magnetic Materials Database provides a large array of datasets for magnetic compounds as well as magnetic clusters, with focus on rare-earth-free magnets. Available datasets include (i) crystallographic data, (ii) thermodynamic properties, and (iii) magnetic properties.	https://www.novomag.physics.ias.tate.edu/structure-database
AtomWork-Adv	Inorganic materials database, including crystal structure, material property, phase diagram, publication, X-ray diffraction.	https://atomwork-adv.nims.go.jp/
Bradley Melting Point Database	Jean-Claude Bradley's Legacy Dataset of Open Melting Points, including 28,645 measurements.	https://figshare.com/articles/dataset/Jean_Claude_Bradley_Open_Melting_Point_Datset/1031637
FeeSolv	The Free Solvation (FreeSolv) database provides experimental and calculated hydration free energies for small molecules in water	https://github.com/MobleyLab/FreeSolv
PubChem	PubChem is the world's largest collection of freely accessible chemical information. Search chemicals by name, molecular formula, structure, and other identifiers. Find chemical and physical properties, biological activities, safety and toxicity information, patents,	https://pubchem.ncbi.nlm.nih.gov/

	literature citations and more.	
Pauling file	The Pauling file collects inorganic material data from more than 35,000 publications from 1900 to the present, containing 350,000 crystal structures, 50,000 phase diagrams, and 150,000 physical properties	https://www.paulingfile.com/
PoLyInfo	PoLyInfo provides information on hundreds of chemical structures, properties, and synthetic methods of polymer materials	https://polymer.nims.go.jp/
Materials Cloud	Materials Cloud is an open materials science platform designed to enable seamless sharing and dissemination of computational materials science resources, providing education, research, simulation software, and curated raw data.	https://www.materialscloud.org

Supplementary table 3: Toolkits of high-throughput computation in materials science

Program	Description	DFT software interface
MatCloud	MatCloud is a platform integrating high-throughput calculations and materials databases	VASP
AFLOW	AFLOW is an automated high-throughput computational materials discovery framework developed by Duke University, which can be used to perform high-throughput calculations on the properties of alloys, inorganic crystal structures, and other materials.	VASP, QE
FireWorks	FireWorks is an open-source software for high-throughput calculation tasks, which has completed millions of high-throughput calculation tasks in computational chemistry and materials	VASP
JARVIS	JARVIS is a data-driven material design toolkit that integrates DFT computational properties of materials, force field calculations databases, machine learning	VASP, QE, BolzTrap, LAMMPS, WIEN2K

	predicted properties databases and related high-throughput calculations and machine learning tools.	
ASE	ASE is a software package written in the Python language to easily calculate the properties of materials and other high-throughput computational simulation tasks.	VASP, QE, CASTEP
AiiDA	AiiDA is an open source high-throughput computing framework that integrates the functions of automation, management, storage, sharing, and reproduction of computing and related data tasks.	QE
SEHC	SEHC is a high-throughput computing software package featuring self-assessment screening. SEHC could improve the efficiency of high-throughput calculations by evaluating the rationality of high-throughput calculation tasks and interrupting unreasonable high-throughput computing tasks in advance.	VASP
JAMIP	JAMIP is a large-scale computational material design software package that integrates high-throughput computing, data generation, collection, management and storage, and machine learning.	VASP
Pymatgen	Pymatgen provides application programming interfaces to a large number of first-principles DFT calculation software such as VASP, ABINIT and Gaussian to support materials files in various formats	VASP, ABINIT, Gaussian
RadonPy	RadonPy, an open source Python library for fully automated polymer property calculations for all-atom classical molecular dynamics simulations, has successfully performed high-throughput calculations on more than 1000 amorphous polymers with a wide range of thermophysical properties	LAMMPS

Supplementary table 4: Commonly used high-throughput preparation and characterization techniques

method

Preparation/

description

Diffusion multi-node method	Preparation	Diffusion multi-node method is a high-throughput preparation method for forming binary and ternary diffusion nodes by using diffusion within a certain composition range to form multiple bulk metals arranged in a predetermined manner with continuously changing composition at a certain temperature.
Co-deposition	Preparation	The co-deposition method utilizes the change of the relative angle and position of each deposition source and the substrate to deposit a variety of alloy components on the same substrate to form a material sample with a gradient distribution of alloy components.
Continuous mask method	Preparation	The continuous mask method is to control the deposition rate of the sample by controlling the substrate stage, and then use the coating and time-shifting mask technology to form a composite sample of multi-component alloys with controllable composition
Discrete mask method	Preparation	The discrete mask method is a method of combining coating and mask technology, and using continuous masks to obtain samples of different discrete components, which is suitable for the preparation of new materials with multiple components and a large component space span.
Spray synthesis	Preparation	The spray synthesis method is to deposit raw materials of different components in a substrate or a reaction chamber by spray technology, so as to obtain a multi-component composite material sample.
Microfluidics	Preparation	Microfluidic technology is a preparation technology

that uses microchannels to precisely control microscale fluids. The main platform of this technology is a microfluidic chip, which can be integrated into basic operations such as sample preparation and characterization.

The material gene combinatorial chip technology is to integrate discrete samples into a sample library or reaction library in an array, and use the ideas and theories of combinatorial theory to quickly and efficiently prepare a large number of samples through various chemical reactions of various reactants or precursor compounds.

Due to the different relative atomic masses of different elements, under the action of strong centrifugation and heating, the centrifugal force on each component of the material is different, so the strong gravitational field can be used to gradually convert a homogeneous material into a material with a composition gradient

Synchrotron radiation sources enable high-brightness microfocusing across the full spectral range, increasing the luminous flux density, brightness, and resolution required for high-throughput characterization

The spallation neutron source is equipped with a spectrometer that can measure the energy and dynamics of high-throughput samples to accurately capture the overall properties of a substance

The digital three-dimensional microstructure characterization is to obtain a two-dimensional image through mechanical polishing technology,

Materials gene combination
chip technology

Preparation

Strong gravitational field
sedimentation atomic method

Preparation

Synchrotron radiation source

Characterization

Spallation neutron source

Characterization

Digital 3D microstructural
characterization

Characterization

		and then use a computer to process the two-dimensional image into a digital three-dimensional structure.
Time domain heat reflection technology	Characterization	Time-domain thermal reflectance technology can quickly measure the time-dependent relationship of laser reflectivity and the time-dependent relationship of sample surface temperature.
Nanoindentation	Characterization	Nanoindenter can characterize the strength, hardness, elastic modulus and other mechanical properties of nanoscale materials.
Scanning probe microscope	Characterization	Scanning Probe Microscopy is suitable for qualitative analysis and characterization of high-throughput material structure and surface micro-regions, producing high-resolution images
Evanescent microwave probe microscopy	Characterization	Evanescent microwave probe microscopy is used for high-throughput characterization of the electromagnetic properties of material chips due to its very high micro-area resolution and precise data acquisition system.
Parallel nano-scanning calorimeter	Characterization	The Parallel nano-scanning calorimeter device consists of a series of substrates of micro heating or calorimetry units, which can simultaneously measure the thermodynamic parameters such as enthalpy change, heat capacity, phase transition temperature, etc.

Supplementary table 5: Commonly used imbalanced learning methods

method	technical foundation	description
RUS	Sampling	Randomly sample the majority class samples and remove them from the dataset until the desired class distribution is reached
RUSBoost	RUS/	RUSBoost uses a random undersampling technique to remove

	Boosting	instances from the majority class in each iteration of AdaBoost.M2, normalizing instance weights in the new undersampled dataset
NUS	KNN	NUS combines undersampling techniques with noise filtering to filter out noisy data from the original minority dataset, and then train a classifier using the new numerical dataset
DBU	KNN	DBU selects a certain number of samples as resampling majority class data according to the similarity coefficient of samples. For the minority class samples, the noise is eliminated by deleting the examples with the similarity coefficient of 0, so as to achieve the purpose of sampling the data and adjusting the imbalance phenomenon.
Tomek Links	Sampling	Two samples of different classes are defined as a Tomek Link pair, and the noise data is removed by the third sample point distance constraint
NCL	Sampling	NCL randomly finds a certain number of neighbor samples, and deletes a certain number of majority classes by judging the sample categories, while the minority classes are not processed.
SMOTE	SMOTE	SMOTE generates a new minority class sample without repetition by randomly selecting the samples of the same kind of neighbors for interpolation
SMOTEENN	SMOTE	SMOTEENN uses SMOTE to generate new minority samples, and then uses the ENN algorithm to eliminate Tomek Links pairs in the new data set after obtaining the expanded data set.
KM-SMOTE	SMOTE/ K-means	KM-SMOTE first uses the K-means algorithm to cluster the minority class samples, and uses the obtained clusters as regions to perform SMOTE interpolation
Borderline-SMOTE	SMOTE	Borderline-SMOTE Same oversampling technique as SMOTE, only oversampling the boundaries of the minority class
Safe-Level-SMOTE	SMOTE	Safe-Level-SMOTE carefully samples minority instances with varying degrees of weight, assigning each positive instance a safety level before generating instances
DBSMOTE	SMOTE	DBSMOTE generates samples along the shortest path from each

		positive instance to the pseudo-centroid of the minority class cluster
SMOTEBagging	SMOTE/ Bagging	SMOTEBagging combines the SMOTE algorithm into the bagging model and extends it to solve multi-class datasets with improved overall performance and diversity
SMOTEBoost	SMOTE/ Boosting	SMOTEBoost introduces SMOTE in each iteration of Boosting to learn a wider minority class decision region
AEIRF	RF	Through the combination of mixed sampling strategy and random forest, AEIRF not only takes advantage of the advantages that mixed sampling can process the data set more comprehensively, but also combines the advantages of random forest to balance errors and strong anti-overfitting ability.
ARIRF	RF	ARIRF improves the diversity of the base classifier and enhances the effect of the classifier by adopting a mixed sampling strategy for each subtree in the random forest
NOBDF	SVM	NOBDF combines oversampling and SVM for dataset reconstruction, and has achieved good results in both positive and negative classification accuracy and overall classification performance.
FSVMs	SVM	FSVMs combine the oversampling method with the fuzzy semi-supervised SVM learning method to further solve the multi-class imbalance problem based on the segmentation strategy
GAN	DL	The loss function and network structure of GAN are appropriately modified using WGAN (Wasserstein GAN) to make it more stable in training. Compared with SMOTE for data augmentation experiments, WGAN is less sensitive to thresholds than SMOTE.
GAN-DAE	DL	GAN-DAE improves the imbalance of data samples by acquiring the characteristics of positive and negative samples in imbalanced data through the confrontational training of generator and discriminator.
DNN	DL	DNN extracts the features of a few samples as basic features, and then adds some pseudo-features to generate new samples to make up for the shortage of minority samples, effectively improving the classification results of imbalanced datasets

D -WELM	ELM	D-WELM not only considers the influence of the number of sample categories, but also considers the influence of the data distribution characteristics, that is, the degree of data dispersion
CHMDT	DT	CHMDT is used for binary classification of unbalanced datasets, and the two classes are unbalanced to improve the classification accuracy
WBCRF	K-means	WBCRF uses the K-means clustering method for undersampling, and selects the attribute with the largest drop in the misclassification cost for division.
