Computational Materials Science 117 (2016) 270–278

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries

M. Attarian Shandiz *, R. Gauvin

Department of Materials Engineering, McGill University, Montreal, Quebec H3A 0C5, Canada

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 8 February 2016
Accepted 11 February 2016

Keywords:
Li-ion batteries
Crystal system
Machine learning
Classification methods
Monte Carlo
Ensemble methods
Optimization

ABSTRACT

The system of crystal structure has a major effect on the physical and chemical properties of Li-ion silicate cathodes. Hence, the prediction of crystal system has a vital importance to estimate many other properties of cathodes for applications in batteries. Three major crystal systems (monoclinic, orthorhombic and triclinic) of silicate-based cathodes with Li–Si–(Mn, Fe, Co)–O compositions were predicted using wide range of classification algorithms in machine learning. The calculations are based on the results of density functional theory calculations from Materials Project. The strong correlation between the crystal system and other physical properties of the cathodes was confirmed based on the feature evaluation in the statistical models. In addition, the parameters of various classification methods were optimized to obtain the best accuracy of prediction. Ensemble methods including random forests and extremely randomized trees provided the highest accuracy of prediction among other classification methods in the Monte Carlo cross validation tests.

1. Introduction

The advancements in numerical methods to calculate electronic structure of materials besides the rapid improvements in the computational power have provided the opportunity of computing physical and chemical properties of a wide range of novel and complex materials [1–3]. Consequently, researchers have access to enormous amount of information about the estimated properties of materials. As an example, Materials Project [4–6] offers an open web-based access to the calculated physical and chemical properties of known and predicted materials derived from density functional theory (DFT) calculations of electronic structure. DFT calculations are powerful methods for the estimation of electron density and band structure of materials. The progression in development of exchange–correlation potential has led to many precise computations of physical properties for many diverse types of materials including Li-ion batteries [7–10]. Subsequently, the huge amount of information about materials should be analyzed to achieve an improved understanding of materials properties. Generally, the complex correlations between different physical properties are hard to discover using traditional statistical models. However, advanced machine learning (ML) methods have the potential to discover the complex correlation between crystal structure and different physical and chemical properties. ML has been used for solving many complex classification and regression problems in numerous scientific felids such as prediction of physical properties [11], corrosion rate [12], lattice parameter [13], crystal structure [14,15], 3D reconstruction of cells in microscopy [16], and many applications for Li-ion batteries [17–20].

Cathode materials with Li–Si–(Mn, Fe, Co)–O compositions are in great interest for research due to their applications in Li-ion batteries. For example, compounds with orthosilicate structure (Li_2XSiO_4, X = Mn, Fe, Co) are one of the major candidates as suitable cathodes for Li-ion batteries because of their low production cost and providing high capacity and safety [21,22]. Crystal structure of cathodes have a significant effect on the properties of Li-ion batteries [23]. Therefore, investigation and development of suitable computational and experimental methods for the characterization of cathodes are fundamental for the better understanding of their physical and chemical properties.

In this research, various classification algorithms are investigated to predict the three major types of crystal system (CS) (monoclinic, orthorhombic and triclinic) of cathode materials with Li–Si–(Mn, Fe, Co)–O compositions using the data from Materials Projects. The majority of DFT results for predicted or known cathodes are available for these three classes. The ML methods to build the models are linear, quadratic and shrinkage discriminant analysis, neural networks, support vector machines, k-nearest neighbors, random forests and extremely randomized trees. The performance

* Corresponding author.
E-mail address: mohammad.attarianshandiz@mail.mcgill.ca (M. Attarian Shandiz).

http://dx.doi.org/10.1016/j.commatsci.2016.02.021
0927-0256/

of classification methods are evaluated based on Monte Carlo cross validation tests on the dataset.

It should be emphasized that the features (properties) are dependent on the crystal structure as the main input for DFT calculations. Hence, the correlation between the predicted values for features and CS is anticipated. However, the main goal of the presented approach is to answer these questions: (1) is it possible to predict the CS having other materials properties? (2) what features are more important for this prediction? The answer to the first question is positive; although, the prediction can be achieved using proper statistical learning methods as described in this paper. The presented approach in this study can be useful for other researchers to consider the correlations between features in the results derived from high performance calculations. In fact, this type of investigation can lead to a better insight regarding the relationship between various features of materials.

2. The dataset

The dataset contains the results of DFT calculations for 339 cathode materials with Li–Si–(Mn, Fe, Co)–O compositions using the data from Materials Project. In Materials Project [4–6], the DFT calculations and optimizations are performed using VASP software [24]. The exchange–correlation potentials for DFT calculations in Materials Project are generalized gradient approximation (GGA) or GGA + U [4]. Materials Project is based on a high-throughput process. Many of the crystal structures for DFT calculations in Materials Project are from inorganic crystal structural database (ICSD) containing positions of atoms and lattice parameters of crystals [6]. The optimization of atomic positions are also performed on available or generated structures. The initial DFT calculations can be based on available data from ICSD, previous calculations, modified structure by chemical substitution and contributions from user community of the project [4]. More information about the details of calculations can be found in the paper by Jain et al. [4].

The dataset contains the chemical formula, space group, formation energy (E_f), energy above hull (E_H), band gap (E_g), number of sites (N_s), density (ρ), volume of unit cell (V) and CS of each cathode. The aforementioned properties in the dataset can be defined according to the glossary of Materials Project as follows. N_s and ρ are the number of atoms in the unit cell of crystal and the density of bulk crystalline materials, respectively. To build ML models only variable V is used given that $V = M/\rho$ (M is the atomic mass). Also, E_H is defined as the energy of decomposition of material into the most stable ones [6]. It should be noticed, the calculation of formation energy and other properties are at the temperature of 0 K and ambient pressure. E_g and V can be dependent on temperature and pressure of system; however, for our calculations the temperature and pressure are considered constant. Table 1 shows the data for some selected silicate cathodes from the dataset. The dataset contains a wide range of complex structures and various chemical compositions.

Fig. 1 shows the pair plots of the properties of silicate cathodes in the dataset. The diagonal plots are the histogram plots for the distribution of each feature of cathodes. As it can be seen, generally there is no evident correlation between the features and the CSs. This complexity makes the classification problem hard to be solved by conventional methods. It should be mentioned the results of calculations in the Materials Project are prone to change because of performing new optimizations or using novel potentials.

3. Methods of classification for machine learning

Classification is a method in ML to split the dataset into certain classes. Since the CSs (monoclinic, orthorhombic and triclinic) are specified, the ML is called a supervised learning. Also, the accuracy of classification is defined as the portion of correct prediction of classes. The feature matrix, X, with $n \times m$ dimensions and the response matrix, Y, as a one dimensional matrix with length n and K different classes are used for the supervised classification. Here n is the number of observations (samples) and m is the number of features. For this study n, m and K are 339, 5 and 3, respectively. CS can be defined as a function depending on other variables as: $CS = f(V, E_g, N_s, E_f, E_H)$. In fact, based on five variables of V, E_g, N_s, E_f and E_H the class of CS can be estimated using ML methods. In this section, the applied classification methods on the dataset to build the models are concisely introduced. The mathematical details of applied methods can be found in the cited papers.

3.1. Linear, quadratic and shrinkage discriminant analysis

Linear discriminant analysis (LDA) is based on the estimation of the distribution of predictors (X) in the response classes, *i.e.* $f_k(X) \equiv \Pr(X = \boldsymbol{x}|Y = k)$ where $f_k(X)$ is the density function of X for the class k [25]. Afterward, using Bayes' theorem the probabilities of occurring the response in each class ($\Pr(Y = k|X = \boldsymbol{x})$) are calculated. So LDA based on Bayes' theorem can be formulated as [25]:

$$\Pr(Y = k|X = \boldsymbol{x}) = \pi_k f_k(\boldsymbol{x}) \Big/ \sum_{l=1}^{K} \pi_l f_l(\boldsymbol{x}) \tag{1}$$

where π_k is the prior probability of class k. LDA uses normal distribution for estimation of f_k and assumes the covariance matrix is the same for each class [26]. In contrast to LDA, quadratic discriminant analysis (QDA) presumes each class can have different covariance matrix leading to possibly a better classification accuracy [26].

Shrinkage discriminant analysis (SDA) is based on LDA or diagonal discriminant analysis (DDA) [27]. DDA is an special case of LDA when covariance matrix is diagonal [28]. In fact, LDA and DDA act as the ranking predictors and SDA uses feature selection for the enhancement of accuracy of classification [27,28]. The sda

Table 1
Data for some selected silicate cathodes from the dataset.

Formula	Space group	E_f (eV)	E_H (eV)	E_g (eV)	N_s	ρ (g/cm^3)	V (Å^3)	CS
Li_2MnSiO_4	Pc	−2.699	0.006	3.462	16	2.993	178.513	Monoclinic
$Li_2Mn_2(SiO_3)_3$	P21/c	−2.769	0.077	3.188	64	2.517	929.064	Monoclinic
$Li_2Co_2(SiO_3)_3$	P21/c	−2.598	0.069	2.727	64	2.739	872.856	Monoclinic
$Li_2FeSi_3O_8$	P21	−2.84	0.069	3.081	28	2.665	351.384	Monoclinic
$LiMn(SiO_3)_2$	Pbca	−2.824	0.036	0.037	80	3.343	850.626	Orthorhombic
$LiFeSiO_4$	Pn21a	−2.604	0.018	2.961	28	2.89	355.979	Orthorhombic
$Li_2Co_2Si_2O_7$	C2cm	−2.453	0.072	2.84	26	3.579	278.304	Orthorhombic
$Li_7Mn_{11}(Si_3O_{16})_2$	P1	−2.439	0.092	0.361	56	3.909	566.407	Triclinic
$LiFeSi_3O_8$	P1	−2.896	0.032	3.342	26	2.76	330.953	Triclinic
$LiCo_3(SiO_4)_2$	P1	−2.25	0.076	0.005	42	3.318	552.402	Triclinic