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/© 2016 Elsevier B.V. All rights reserved.

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]:

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

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

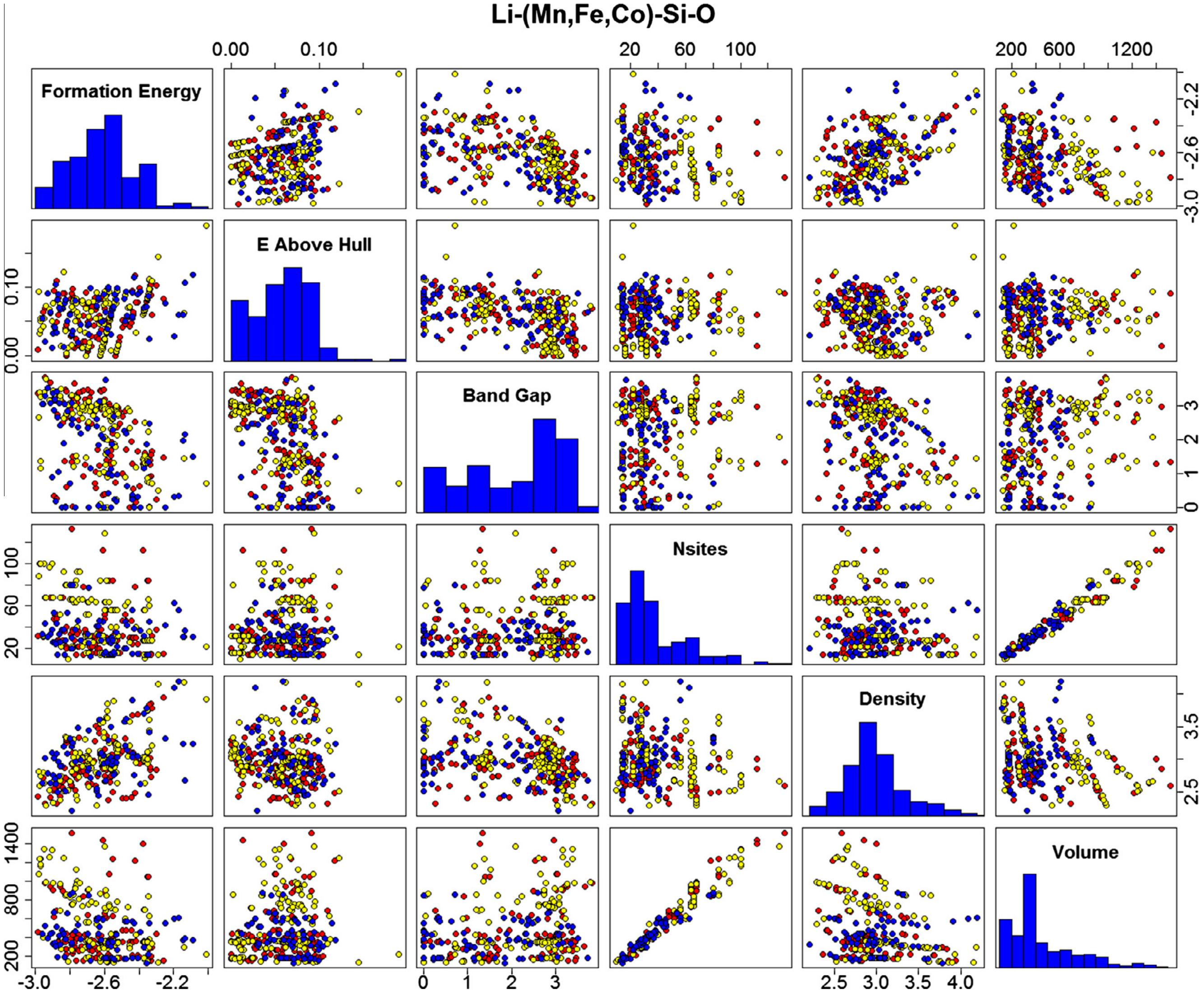

Fig. 1. The pairs plot of different properties of Li–(Mn, Fe, Co)–Si–O cathodes based on the extracted data from Materials Project. The red, yellow and blue circles indicate the monoclinic, orthorhombic and triclinic crystal systems respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

library in R [28] was used is based on James-Stein-type shrinkage estimators for correlation matrix [29] and variances [30]. So, the shrinkage intensities for SDA are correlation matrix and variances. The value of shrinkage parameter can change between 0 and 1 and the LDA classifier was used for SDA.

3.2. Artificial neural networks

Artificial neural networks (ANN) are one of the most versatile and effective methods in ML for classification and regression problems [31]. However, the high degree of flexibility can make the process of finding the optimal values of parameters difficult. In this study for the simplicity in the approach, feed-forward neural networks with a single hidden layer [32] was used. Feed-forward neural networks have three different type of layers including input, hidden and output layers; though, the number of hidden layer is limited to one [33,34]. For our classification, the main parameter for optimization was number of unites in the hidden layer. The activation function was considered the logistic sigmoid function and maximum conditional likelihood was used as fitting criterion [32].

3.3. Support vector machines

Support vector machine (SVM) methods work on the idea of splitting the training data based on the finding the hyper-planes with the maximum margin. The hyper-plane is defined as [35]: $\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b$, where $\boldsymbol{w}$ is the normal vector to the hyper-plane, $\boldsymbol{x}_i$ is the training dataset and $\phi(\boldsymbol{x}_i)$ maps the training data to the feature space. To find the optimal solution of maximum margin, the following optimization problem should be solved [36]:

$$\begin{aligned} &\min(\boldsymbol{w}, b, \xi): \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to}: \ y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b) \geqslant (1-\xi_i); \ \xi_i \geqslant 0 \end{aligned} \tag{2}$$

where N is the number of training data, $y_i \in \{-1, 1\}$ indicating the positive and negative values for separation and $C > 0$ is a regularization parameter as a cost evaluation that puts penalty based on the amount of the training error and the complexity of model [35]. Also, ξ_i is used to obtain the soft margin of the classifier.

The hyper-planes can be linear or non-linear. The non-linearity of hyper-planes is introduced to the SVM classifier using kernel function, K, and Lagrange multiplier, α_i, as below [35,37]:

$$\begin{aligned} &\max: \ \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{subject to}: \ \sum_i \alpha_i y_i = 0; \ 0 \leqslant \alpha_i \leqslant C, \ i = 1, \ldots, N \end{aligned} \tag{3}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. For this study, the radial basis function was used as the kernel function with the following formula [38]:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \tag{4}$$

where γ is a regularization factor that controls the kernel function. Finally, the classifier can be obtained as follows [36]:

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \tag{5}$$

It should be noticed that the data for training are scaled to have mean and variance equal to zero and one, respectively.

3.4. k-nearest neighbors

k-nearest neighbors (kNN) is a simple and effective method that can be used for classification. kNN resolves the class of a new observation based on finding the closest observation in the dataset by evaluation of similarity using the measurement of distance [25,39]. Hence, the number of nearest neighbors and the method for measuring distance are the two key parameters in kNN method. For this study, the distance was measured using Minkowski distance as below [40]:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{s=1}^{\delta} |\boldsymbol{x}_{is} - \boldsymbol{x}_{js}|^p\right)^{\frac{1}{p}} \tag{6}$$

where δ is the size of vector $\boldsymbol{x}$. Euclidean and absolute distance with $p = 2$ and $p = 1$ are two special cases of Minkowski distance. For the calculations, the optimal value of p was determined in Monte Carlo cross validation tests using weighted kNN classifier. In weighted kNN method, after standardization of distances, kernel function is used to transform distances to weights [40]. The details of weighted kNN can be found in the paper by Hechenbichler and Schliep [40]. The inversion kernel $(1/|d|)$ was used as the kernel function for kNN classification.

3.5. Random forests and extremely randomized trees

Random forests (RF) and extremely randomized trees (ERT) are both ensemble ML methods based on building decision trees. RF is based on bagging method which makes a variety of decision trees by means of bootstrap sampling from the training data and averaging to build the total decision tree [41]. The classification is based on the majority vote after averaging. However, to improve the process of decorrelation of variables and increasing the chance of building better decision trees, variables are selected randomly at each split [42]. The random selection of variables decreases the variance and improves the overall accuracy of prediction. The default number of randomly selected values (m_v) for RF is equal to the square root of number of predictors in the training data [25]. Although, this number can be different and $m_v = 3$ was selected for this study due to providing the highest accuracy. Similar to RF, ERT [43] uses the random selection of predictors at each node. Though, the major dissimilarity is that RF utilizes the best cutting (discretization) threshold but ERT selects this cut randomly with values not smaller than a specified threshold [44,45]. This extra level of randomness can improve the process of decorrelation and can lead to better accuracy of prediction by the classifier. In addition, the whole training dataset is used for making various decision trees in ERT instead of using bootstrap sampling in RF [46].

4. Results and discussions

Building ML models and calculations were performed using R language and the related packages for classification [47]. The following libraries were used for calculations: MASS for LDA and QDA, sda for SDA, nnet for ANN, e1071 for SVM, kknn for kNN, randomForest for RF and extraTrees for ERT. The details about the libraries can be found at Ref. [47].

Monte Carlo cross validation (MCCV) or called repeated random sub-sampling method was used for the evaluation of the accuracy of models. MCCV has been used by many other researchers for estimation of the prediction error and selecting best models [48–51]. MCCV is based on dividing dataset into two subset of training and testing randomly [50]. The process is repeated and for each generated sub-sample the accuracy of prediction by classifier is calculated. A good estimation of the overall accuracy of prediction can be computed by having large enough number of repetitions in MCCV tests [50].

It is worth to mention for a true random sampling, each time first the dataset was shuffled randomly and then it was divided into two subsets. MCCV has the benefit of being asymptotically consistent [48–50] and have a good chance of choosing the best model with precise prediction of accuracy [51]. It should be noticed for generation of 100 random samples 100 random seeds corresponding to 100 states of random number generator were used. Hence, the 100 samples are identical for all the classifiers providing the ability of one-to-one comparison of the results.

Fig. 2 demonstrates effect of percentage of training data on (a) the overall average accuracy (OAA) of prediction and (b) the standard deviation (SD) of OAA in MCCV tests using all the 8 ML methods. OAA is the average accuracy of classification for each ML method based on MCCV for 100 random samples. As it can be seen, increasing the percentage of training data continuously improves OAA of prediction and increases the SD. RF and ERT gave the highest accuracy with ~75% and ~76% respectively using 90% of data for training the model. The increase of SD of OAA is expected because decreasing the number of testing data leads to the increase of SD. However, methods with highest OAA such as RF and ERT show relatively low SD at high percentage of training data. This can be a result of flexibility of RF and ERT as ensemble methods to obtain the highest accuracy for each random sample and reduce the SD of prediction. Increase of the percentage of training data does not improve the accuracy of prediction for LDA and QDA confirming the two methods are not appropriate for the classification of CS in this study. This means the normal distribution assumption is not valid for the estimation of the distribution of variables in the response classes. More details about the optimized parameters of classifiers are discussed in Figs. 4–10.

Since MCCV was used for the testing the accuracy of ML models, the effect of random number generator (RNG) is also important to consider due to the stochastic nature of the process. To examine the effect of RNGs, 6 RNG methods were used to generate random seeds. Random seeds are vectors of numbers that can estimate the properties of a random sample. The random seeds (pseudo-random numbers) are predefined as different states and they can be identified by their state. The process of using specific random seeds is essential for many stochastic simulations because the results can be reproduced. The RNG used to generate the random seeds are

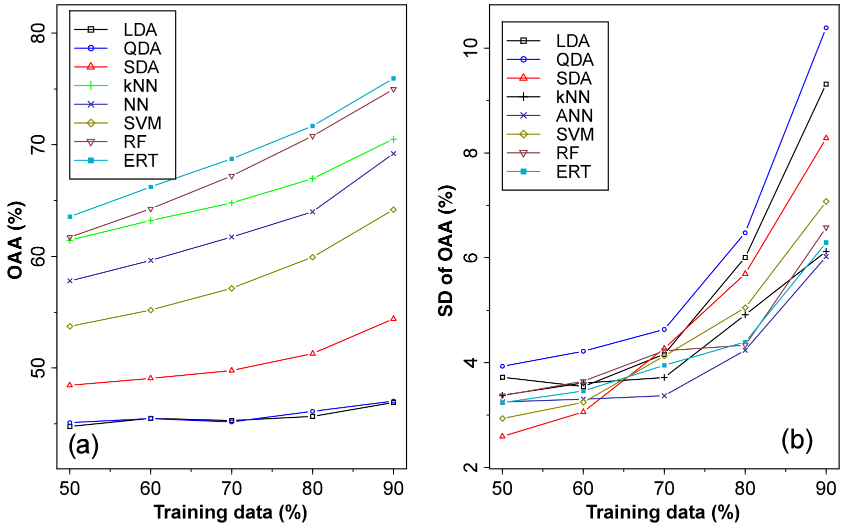

Fig. 2. Effect of percentage of training data for building ML models on (a) the overall average accuracy (OAA) and (b) SD of OAA using 100 random samples in MCCV tests.

Wichmann–Hill, Marsaglia-Multicarry, Super-Duper, Mersenne-Twister, Knuth-TAOCP-2002 and L'Ecuyer-CMRG from random seeds in R library [47]. All the 6 RNG have very large cycle length and there is a very low probability for having the same random samples. More details about the algorithms and their cycle length can be found at [47].

Table 2 and Fig. 3 show the results of different RNGs on the OAA of 100 random samples. The results are based of 80% of data for training the model and 20% for testing. Generally for a certain ML method the average accuracies are very close and the SDs are small. This confirms by using proper RNGs the average accuracies of models can be reproduced and the accuracy of models is mainly related to the algorithms not the random sampling process. It should be noticed that Mersenne-Twister was used as the RNG for the random seeds for all the other results represented in Figs. 2 and 4–10.

Fig. 4 illustrates the histogram and density plots of the accuracy of different applied classification methods. The results in Fig. 4 are based on using 90% of each random sample for training and 10% for validation of the model. As it can be seen, LDA, QDA and SDA did not provide a good accuracy of prediction. In some cases LDA and QDA have the prediction even less than the average null prediction (44.9%). This means discriminant analysis method does not work very well for the estimation of CS in the dataset. Although, using SDA improves the average accuracy about 7% in comparison to LDA and QDA. ANN and kNN provided accuracies higher than 69% proving the flexibility of the methods. The OAA of SVM is less than the OAA of ANN and kNN. This can be resulted from the fact that optimization of C and γ in the kernel function is intricate since their values can vary in a wide and continues range. The ensemble methods (RF and ERT) gave the highest accuracy. In addition, only ERT granted predictions with accuracies higher than 90%. The large variation of OAA emphasizes the importance of selecting proper methods for the classification. In addition, the optimization of the parameters of classifiers has a significant effect on the overall accuracy. For example without optimization of parameters, we have seen 10–20% decrease in OAA in MCCV tests. The optimized parameters for different ML methods are presented in Figs. 5–10 using 90% of data for the training of models.

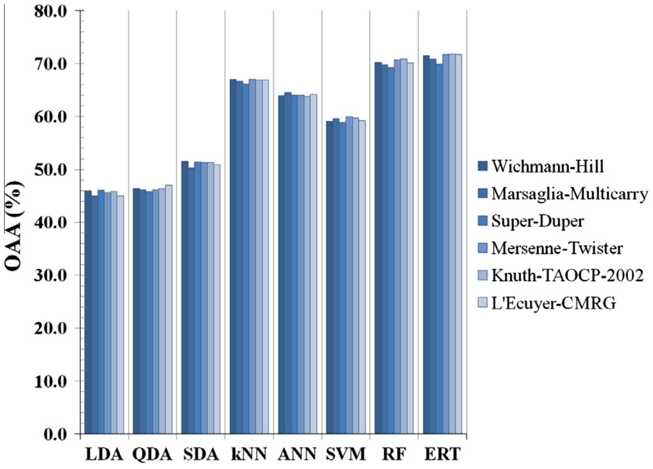

Fig. 3. Effect of different RNGs on the OAA of ML methods based on results presented in Table 2.

Fig. 5 shows the histogram plots of best shrinkage values for the correlation matrix and variances in SDA method for the different

Table 2
Effect of different RNGs on OAA (%) of ML methods based on 100 random samples using 80% training data.

	Wichmann–Hill	Marsaglia-Multicarry	Super-Duper	Mersenne-Twister	Knuth-TAOCP-2002	L'Ecuyer-CMRG	SD
LDA	45.9	45.0	46.1	45.7	45.9	45.0	0.45
QDA	46.4	46.2	45.8	46.1	46.4	47.0	0.39
SDA	51.6	50.3	51.5	51.3	51.3	50.9	0.47
kNN	67.0	66.7	66.1	67.0	66.9	66.9	0.32
ANN	63.9	64.6	64.0	64.0	63.9	64.2	0.26
SVM	59.1	59.6	58.9	59.9	59.7	59.2	0.40
RF	70.3	69.9	69.3	70.8	70.9	70.2	0.58
ERT	71.5	70.8	70.0	71.7	71.8	71.7	0.72

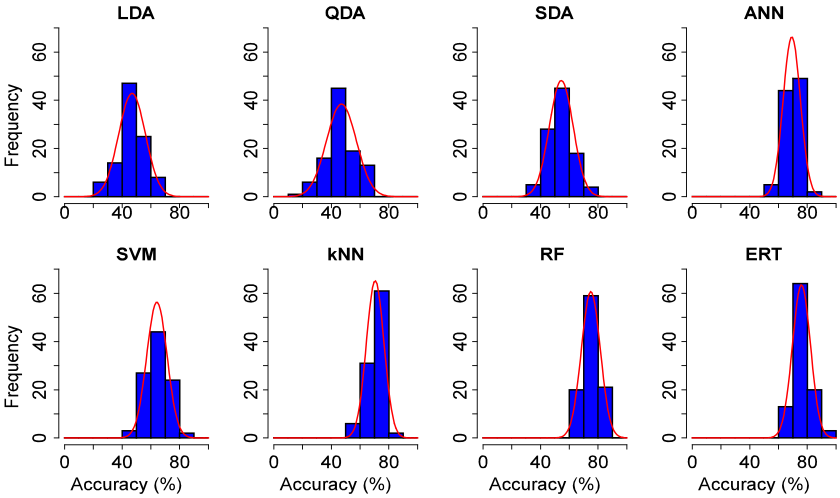

Fig. 4. Histogram and density plots of accuracy of applied classification methods for 100 random samples using 90% training data.

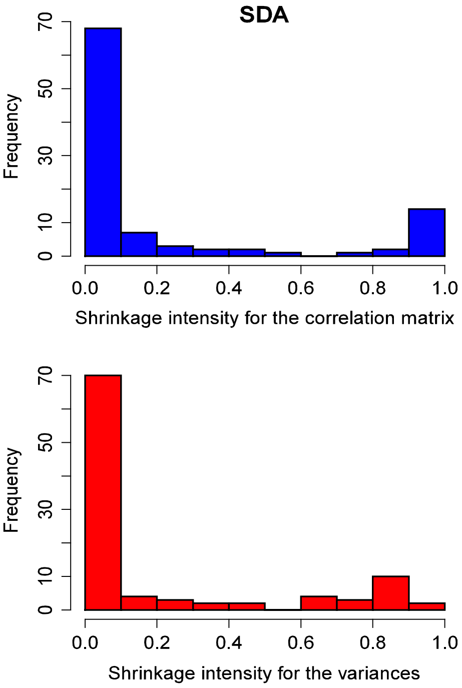

Fig. 5. The histogram plots of best shrinkage intensities for the correlation matrix and the variances in SDA method derived from the results of optimized classifiers for the 100 random samples.

random samples. In contrast to LDA and QDA, SDA has the advantage of more flexibility for tuning the parameters and this can be the reason for about 7.4% higher accuracy in comparison to QDA. As shown in Fig. 5, values of shrinkage for the correlation matrix and variances are dependent on the samples. However, more than 70% of the shrinkage values for the correlation matrix and variances are smaller than 0.2 meaning that small shrinkage provides the best results for the related samples. It should be noted a zero and one shrinkage intensity indicate no shrinkage and a complete shrinkage, respectively [52].

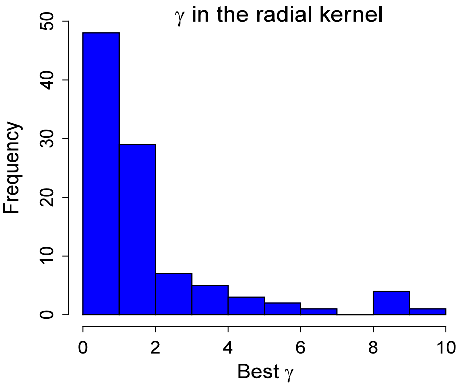

Fig. 6. The distribution of best γ for the radial basis kernel function in SVM for the optimized classifiers.

Fig. 6 presents the histogram plot of optimized values of γ for the radial kernel in SVM method in MCCV tests. The cost value was selected from 1 to 50 with a 5 step length. The optimized value of γ is not the same for different samples and the optimization is essential for achieving better accuracy. Optimization of SVM classifiers are computationally expensive because of the flexibility in choosing γ and cost since they can change in a wide and continues range. Setting high value of cost does not guarantee higher accuracy of prediction *i.e.* by changing γ and a different sample the proper value of cost should be selected.

Fig. 7 shows the histogram for the best number of units in the hidden layer of feed-forward neural networks with a single hidden layer for the 100 random samples. The average value of the best number of units is about 33; however, this value varies in a wide

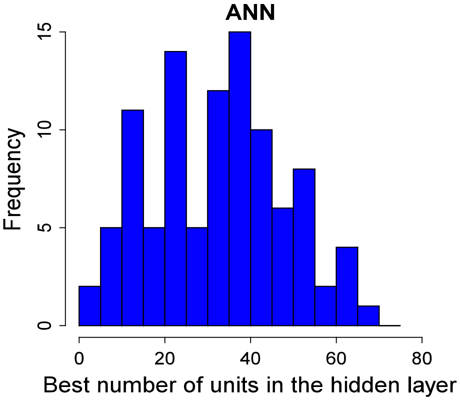

Fig. 7. The distribution of best number of units in the hidden layer in feed-forward neural networks with a single hidden layer.

range. Number of units in the hidden layer was limited to 70 to avoid over complexity of the model for calculations.

Fig. 8a demonstrates effect of parameter of Minkowski distance (p) on the average accuracy of prediction in kNN method. The highest OAA was achieved at about $p = 0.15$. The variation of p changes the distance function in Eq. (6). Hence achieving an optimal value for p means that kNN algorithm provides the highest OAA of prediction based on an optimized kernel function. It should be mentioned each point in Fig. 8a is the results of calculations for all the 100 random samples, *i.e.* for each point MCCV tests have been applied. The histogram plot of best number of neighbors to achieve highest OAA for random samples is shown in Fig. 8b. As it can be seen, the majority of the best number of neighbors are less than 5; although for some cases more than 15 neighbors was necessary to achieve the highest accuracy.

Fig. 9 illustrates the histogram plots of the best number of trees to accomplish the highest OAA in MCCV tests using RF and ERT methods. As shown, using a very big size of tree does not guarantee the highest accuracy. Forcing RF or ERT methods to grow an unnecessary large tree can cause some redundant splits in internal nodes not leading to the maximum reduction in misclassification.

Fig. 10 presents the effect of the percentage of training data on the average number of best size of trees to grow for the highest OAA of predictions in RF and ERT classifiers. Generally, the average number of trees required to grow to achieve the highest OAA decreases by increasing the percentage of training data. This can be explained by the fact that having more data to train for RF and ERT classifiers increases the chance of building better decision trees by proper splits. It should be mentioned for other optimized parameters a significant trend was not observed by variation of the percentage of training data.

Fig. 11 shows the feature importance evaluation of predictors in building ERT models based on the optimized classifiers in MCCV tests using 50% and 90% training data. The decrease in Gini impurity was used as the criterion to evaluate the feature importance of variables. The values of feature importance for ERT were generated using the ExtraTreesClassifier in scikit-learn library [52]. Generally, the difference between the importance of predictors are not significantly different. Hence, it can be inferred all the predictors from the dataset have noteworthy effects for building ERT classifiers. However among other features, the volume of unit cells have the highest importance. The more important role of volume can be related to more sensitivity of CS to the volume distribution. For example, the average volume for monoclinic, orthorhombic and triclinic in the dataset are 438.9, 538.1 and 398.5 (Å^3) respectively. Although, the maximum volume for monoclinic, orthorhombic and triclinic are 1518.9, 1374.7, and 878.3 respectively. In addition, generally by increasing volume N_s increases. Therefore N_s is also an important feature in building ERT classifiers. Variation of percentage of training data does not change the importance of features considerably as shown in Fig. 11.

In theory all the applied ML methods in this research can be used for the multiclass classification problems. However, the underlying assumptions and the flexibility of algorithms are the main reasons for the better accuracy of predictions. For example, LDA and QDA as parametric methods use the normal distribution assumption for the estimation of the distribution of variables in

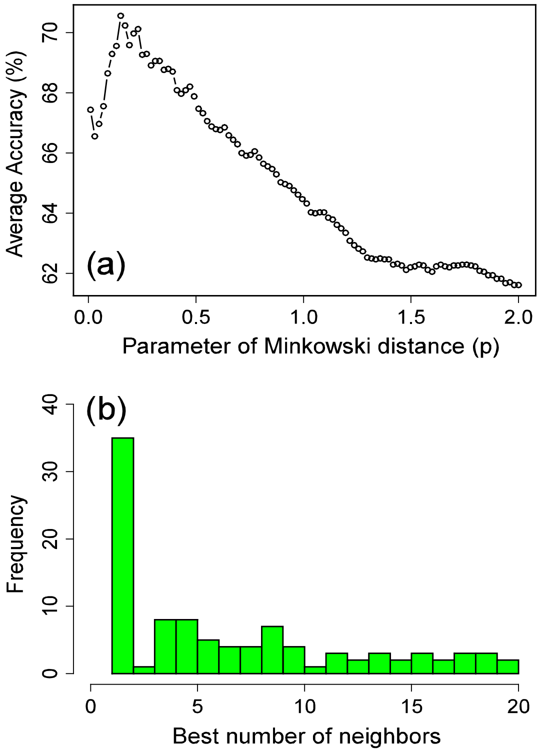

Fig. 8. (a) Effect of parameter of Minkowski distance (p) on OAA of prediction for kNN method. (b) The distribution of best number of neighbors for the optimized kNN classifiers.

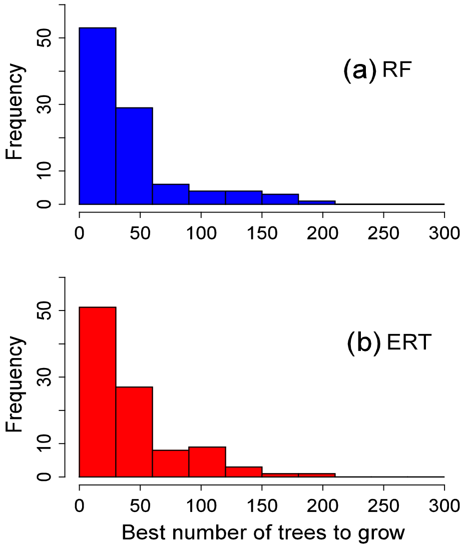

Fig. 9. The distribution of best size of tree to reach highest precision of prediction for the 100 random samples using (a) RF and (b) ERT methods.

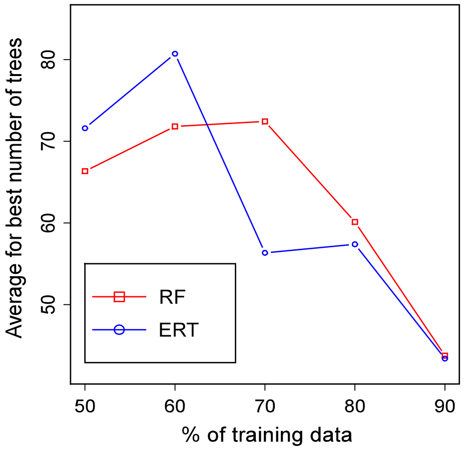

Fig. 10. Effect of the percentage of training data on the average number of best size of trees to achieve highest OAA for RF and ERT classifiers.

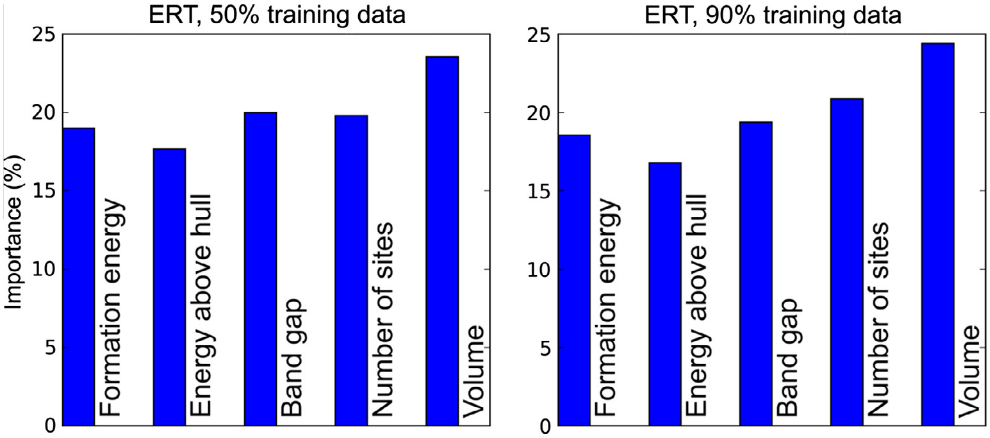

Fig. 11. Feature importance plots of ERT method based on the results of optimized classifiers in MCCV tests using 50% and 90% training data.

the response classes. The validity of applied assumptions plays a key role in the overall accuracy of parametric methods and when the assumptions are not valid the methods fail to achieve a good precision [25]. However, RF and ERT as non-parametric methods are derived from building numerous trees based on the mentioned criterions in Section 3.5. For this reason, the RF and ERT are flexible and there is no need of assumption regarding the distribution of variables in the response class.

It should be emphasized that the size of dataset plays an important role for the improvement of accuracy. It is possible that the accuracy of a method can be improved significantly by having more data. However in this research, increasing the percentage of training data did not change the superiority of a method to the others. Having much more data could change this trend; on the other hand, the number of available data for the silicate cathode materials is limited. The main reason to employ MCCV was to generate much more sub-samples from the limited number of available DFT results to have a better estimation of the real accuracy of ML methods.

5. Conclusions

A wide range of machine learning classification methods were successfully used to determine the three major crystal systems (monoclinic, orthorhombic and triclinic) of silicate cathodes with Li–Si–(Mn, Fe, Co)–O compositions. Monte Carlo cross validation was used for the evaluation of the accuracy of classifiers. It was confirmed that the optimization of parameters of each classification method has a significant effect on the overall average accuracy. Increasing the percentage of training data to build machine learning models continuously increased the overall average accuracy of prediction. Random forests and extremely randomized trees gave the highest overall average accuracy among other classifiers proving the power and flexibility of ensemble methods for the classification of crystal system. Based on the feature importance evaluation in extremely randomized trees, the volume of crystal and number of sites showed the highest effects to determine type of crystal system in the dataset. However, the other features of silicate cathodes including formation energy, energy above hull and band gap are also considerable for determination of crystal system.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.commatsci.2016.02.021.

References

[1] R.F. Service, Science 335 (2012) 1434–1435.
[2] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Chem. Mater. 22 (2010) 3762–3767.
[3] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, Nat. Mater. 12 (2013) 191–201.
[4] A. Jain, G. Hautier, C.J. Moore, S.P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, Comput. Mater. Sci. 50 (2011) 2295–2310.
[5] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Inorg. Chem. 17 (2010) 656–663.
[6] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, APL Mater. 1 (2013) 011002.
[7] L.M. Yan, J.M. Su, C. Sun, B.H. Yue, Adv. Manuf. 2 (2014) 358–368.
[8] Y.S. Meng, M.E.A. Dompablo, Acc. Chem. Res. 46 (2013) 1171–1180.
[9] Y.S. Meng, M.E.A. Dompablo, Energy Environ. Sci. 2 (2009) 589–609.
[10] M. Nishijima, T. Ootani, Y. Kamimura, T. Sueki, S. Esaki, S. Murai, K. Fujita, K. Tanaka, K. Ohira, Y. Koyama, I. Tanaka, Nat. Commun. 5 (2014) 4553.
[11] E. Bélisle, Z. Huang, S. Le Digabel, A.E. Gheribi, Comput. Mater. Sci. 98 (2015) 170–177.
[12] S.F. Fang, M.P. Wang, W.H. Qi, F. Zheng, Comput. Mater. Sci. 44 (2008) 647–655.
[13] S.G. Javed, A. Khan, A. Majid, A.M. Mirza, J. Bashir, Comput. Mater. Sci. 39 (2007) 627–634.
[14] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Nat. Mater. 5 (2006) 641–646.
[15] Ş. Atahan-Evrenk, A. Aspuru-Guzik, Prediction and Calculation of Crystal Structures, Methods and Applications, Springer International Publishing, 2014.
[16] L. Waller, L. Tian, Nature 523 (2015) 416–417.
[17] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C.A.J. Fisher, H. Moriwake, I. Tanaka, Adv. Energy Mater. 3 (2013) 980–985.
[18] J. Du, Z. Liu, Y. Wang, Control Eng. Pract. 26 (2014) 11–19.
[19] C. Hu, G. Jain, C. Schmidt, C. Strief, M. Sullivan, J. Power Sources 289 (2015) 105–113.
[20] Z. Liu, H.X. Li, J. Power Sources 277 (2015) 228–238.
[21] D. Lv, J. Bai, P. Zhang, S. Wu, Y. Li, W. Wen, Z. Jiang, J. Mi, Z. Zhu, Y. Yang, Chem. Mater. 25 (2013) 2014–2020.
[22] H. Lee, S.D. Park, J. Moon, H. Lee, K. Cho, M. Cho, S.Y. Kim, Chem. Mater. 26 (2014) 3896–3899.
[23] R.C. Longo, K. Xiong, K.C. Santosh, K. Cho, Electrochim. Acta 121 (2014) 434–442.
[24] G. Kresse, J. Furthmüller, Comput. Mater. Sci. 6 (1996) 15–50.
[25] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.
[26] R.C. Neath, M.S. Johnson, International Encyclopedia of Education, Discrimination and Classification, 2010, pp. 135–141.
[27] V. Zuber, K. Strimmer, Bioinformatics 25 (2009) 2700–2707.
[28] M. Ahdesmaki, K. Strimmer, Ann. Appl. Stat. 4 (2010) 503–519.
[29] J. Schäfer, K. Strimmer, Statist. Appl. Genet. Mol. Biol. 4 (2005) 32.
[30] R. Opgen-Rhein, K. Strimmer, Statist. Appl. Genet. Mol. Biol. 6 (2007) 9.
[31] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[32] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, 2002.
[33] K. Dai, J. Zhao, F. Cao, Eng. Appl. Artif. Intel. 42 (2015) 57–66.
[34] S. Wang, F.L. Chung, J. Wang, J. Wu, Neurocomputing 149 (2015) 295–307.
[35] S. Kang, P. Kang, T. Ko, S. Cho, S. Rhee, K.S. Yu, Expert Syst. Appl. 42 (2015) 4265–4273.
[36] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2014 <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[37] V.N. Balasubramanian, Conformal Prediction for Reliable Machine Learning, Elsevier, 2014.

[38] P. Wittek, Quantum Machine Learning, What Quantum Computing Means to Data Mining, Elsevier, 2014.
[39] E. Chávez, M. Graff, G. Navarro, E.S. Téllez, Inform. Syst. 51 (2015) 43–61.
[40] K. Hechenbichler, K.P. Schliep, Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Discussion Paper 399, SFB 386, Ludwig-Maximilians University, 2004.
[41] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[42] J. Song, J. Korean Statist. Soc. 44 (2015) 321–326.
[43] P. Geurts, D. Ernst, L. Wehenkel, Mach. Learn. 63 (2006) 3–42.
[44] G. Louppe, P. Geurts, Machine Learning and Knowledge Discovery in Databases, Ensembles on Random Patches, in: ECML/PKDD, 2012, pp. 346–361.
[45] J. Simm, M. Sugiyama, IEICE Trans. Inform. Syst. E97-D (2014) 1677–1681.
[46] L. Rokach, Inform. Fusion 27 (2016) 111–125.
[47] R Development Core Team, R: A Language and Environment for Statistical Computing, The R Foundation for Statistical Computing, Vienna, Austria, 2011 <http://www.R-project.org/>.
[48] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.
[49] Q.S. Xu, Y.Z. Liang, Chemometr. Intell. Lab. 56 (2001) 1–11.
[50] Q.S. Xu, Y.Z. Liang, Y.P. Du, J. Chemometrics 18 (2004) 112–120.
[51] K. Haddad, A. Rahman, M.A. Zaman, S. Shrestha, J. Hydrol. 482 (2013) 119–128.
[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. londel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. 12 (2011) 2825–2830.