Review on Machine Learning Techniques for Crystal Structure Identification

Ravindra Patil[1] Sangeeta Kakarwal[2] and Dhanyakumar Kurmude[3]
[1] Marathwada Institute of Technology, Aurangabad, Maharashtra, 431010, INDIA
[2] P.E.S. College of Engineering, Aurangabad, Maharashtra, 431001, INDIA
[3] Milind College of Science, Aurangabad, Maharashtra, 431001, INDIA
ravindra.be2004@gmail.com

Abstract. Machine learning methods are essential in classification of large data by extracting useful data from it. Machine learning methods works on experienced data to achieve exact identification. Crystal structure identification is one of the most important issues in material science which further leads to assessment of various properties acquired by the materials. Owing to available structural data of crystalline solids, in addition to conventional methods such as X – ray diffraction, Infrared spectroscopy etc. various machine learning approaches can also be employed and the material and one can predict the properties first and then to synthesize the material as desired. Some of these machine learning methods such as Random forest, KRR, DFT, Naive bayes employed by various researchers in structural identification crystalline solids are discussed with its merits and demerits in the monograph.

Keywords: Machine Learning, Crystal Structure Identification, Decision Based Trees, SVM.

1 Introduction

The characterization of small structures or small sized material in the nanometric sales usually calls for sophisticated characterization tools. Different methods of characterization are employed to determine phase, structure and properties of the nanomaterials.

Conventionally, once synthesized, a material should be tested for its structural accuracy or perfection first and then other characterizations for different properties such as electrical, magnetic, physical, chemical, optical be carried out with proper and required means. With the development of new synthesis routes there has been also a great revolution in analytical or characterization tools which made it possible to assess the synthesized material with more accuracy and measure its properties over wide range of different parameters such as temperature, frequency etc.

Though many characterization tools/techniques are available today, very first of the all techniques that come to mind is Bragg's X-ray diffraction for structural analysis of a material. Without having X-ray diffraction at the sample under investigation it is almost difficult to proceed for further studies of the sample in the field of material science. Another powerful technique is IR spectroscopy which also gives much information about the structure and forces building the structures. For surface morphology scanning electron microscopy (SEM) is one of the best and primary mean. EDAX, another technique which is generally associated with SEM is good for elemental/compositional study and hence to check the purity of the sample synthesized.

1.1 Machine Learning

At present large structural data of crystalline solids is available in literature and with various agencies such as ICSD. Further, various properties such as physical, chemical, optical, electrical and magnetic etc. have been studied correlated with its structural investigations by large number of researchers and their data is available in the literature. Obviously, computational methods such as data mining, density functional theory can be employed for structural identification. Furthermore software such as CALYPSO, GASP, GRACE and USPEX can be used for crystal structure studies.

The effective methods are useful for generation, managing and utilizing the important information to reduce the time span of predicting new material using various models. This task is achieved using machine learning approach. Machine learning work on principle of past experiences and information for prediction of targeted result. It already implemented in pattern recognition techniques such as face, fingerprint, Iris recognition and Medical image analysis. Machine learning is also implemented successfully in material science field for detecting the material properties, phase diagram prediction and for identification of crystal structure.

Machine Learning is mainly divided into two branches :supervised and unsupervised learning. In supervised Learning, training data is having both input and related output values. Using this training data , Machine learning algorithm have devised the function which is useful for predicting output data for newly supplied input data. In unsupervised Learning, training data having only input values for which pattern have to be predicted [1].

Machine learning algorithms are useful to extract important information from such large databases to make prediction on past experiences for performance optimization. The machine learning method work on the principle of learning inputs to predict decision. The machine learning approach can be useful on crystal structure to classify the material. In machine learning, computer programs are trained from its past experiences to improve its performance [2][3][4].

1.2 Objectives

One can take the research on structural identification of crystalline solids with following objectives.

- Examine the elements required for prediction problem solution.
- Identification of search plan to use of machine learning technique with mechanical methods.
- Use of binary alloy database for structure correlations mining with the help of graphical models.
- To cross – validated predictions of compound material.
- To develop a structure identification method.

2 Literature Survey

Some of the key results/points to be noted in this regard are as following

Cristopher Carl Fischer have implemented crystal structure prediction using machine learning approach on stable state of system. This research has suggested that wrong selection in structure for material property prediction unable to find fast effect. The vital information of structure is necessary for understanding the recognized material. He has designed a novel strategy with the help of past knowledge material dataset. The inorganic material is having various physical properties like band gaps. Three ingredients necessary for identification of structures are energy model, an effective searching strategy through possible structures and method for entropy evaluation .

Density functional Theory, DFT is an effective technique used for calculating electron energy. Local Density Function, LDA is practical approach of DFT used for identification of properties. Energy difference of any crystal structures is an essential for getting systems stable state which is useful for structural identification along with DFT. The ground state of system acts as an important factor for structure problem. Two different searching methods are discussed for ground states in this research are 1. Coordinate based search. 2. Heuristic rules for structural stability.

Cristopher carl Fischer has worked on Au – Zr, Li – Pt and Ag – Mg alloy system. Accurate energy model is used to achieve the structural stability with DFT. Dataset used in this research is Pauling file dataset [5].

Ghanshyam Pilania and et.al has accelerated the material property prediction by implementing Machine learning methodology. These authors have worked on decision rule with the help of DFT[6].

Shujiang yang & et.al. have implemented machine learning approaches to do the structural identification on zeolite data. ICSD database includes zeolite database having 16 attributes crystal structure. 40 framework types utilized to train the model. This approach gives 95% and above result in structural identification with the help of machine learning approach. The Machine learning method used for doing this

work is Random Forest Method. Some of the useful techniques are discussed in the following lines[7].

With the help of computational material design , crystal structure prediction of material can be done before synthesis by atomic arrangement using potential energy surface.

By Appling machine learning techniques on crystal structure database to predict the targeted structure type in given composition of material. Fischer et. al. worked on such prediction problem model which is known as "Data Mining Structure Predictor"[8]. The extension of this work is related to identification of new ternary oxide compound done by Hautier et.al. with the help of DFT calculations. 355 ternary oxides are identified using this approach which are not in ICSD. They have searched their results in PDF4+ database which is related to diffraction data. They have proved the high success rate for prediction of ternary oxides having 146 structural information which is matched correctly in 140 cases [9]. Machine learning approach produce the accurate prediction of material property with minimum timing.

2.1 Classification

Classification is a supervised machine learning technique in which system must be trained on classified training data. Classification done by supervised machine learning system is known as classifier. The well known example is detection system for spam emails in which given set of emails is marked by spam or not – spam emails. This system learns the characteristics of spam emails due to which upcoming emails can be distinguish as spam or not-spam emails.

Decision tree classifier, support vector machine, rule based classifier and naïve bayes, Random Forest, C4.5 classifiers act as classifiers[4].

2.2 Decision Tree classifier

This technique uses learning algorithm for building predictive model to classify unknown records. It includes series of well defined questions related to attributes of test record. It organizes test questions and condition in tree like structure. The root and internal node include the attribute test conditions for which appropriate answer records are generated to get

classify with specific label class. All terminal nodes assign with Yes or No class label.

The various decision tree algorithms uses greedy approach for optimal decision are ID3,Random Forest, SVM, C4.5, CART, SPRINT[10][11][12].

2.3 ID3 (Iterative Dichotomiser3)

ID3 classification algorithm was developed by J. Ross Quinlan. It derives classes from fixed set of instances. It supports top – down approach for decision making of best attribute selection. The first best attribute selected is assign as root node for decision tree. The decedent node is selected for every possible values associated with root attribute. This downward approach is repeated for training sample dataset to find out best descendent at each point of decision tree.

The inductive classes created for prediction of future classification by ID3 is based on small trained instance of dataset. The sample dataset use for ID3 must have fixed values for each attribute. Every attribute class should be predefined and must have distinct subset of sample dataset. The best Attribute selection is depends on information gain which is used by ID3 for classification of training data into targeted classes.

Entropy is useful for calculating the amount of uncertain information associated with each attribute.

$$\text{Entropy}(s) = \sum_{i=1}^{c} -P_i \log_2(P_i)$$

Where

S is collection of sample dataset having outcome c.

P(i) is proportion of S belongs to class i.

Information Gain

Information gain is essential for reducing the impurity in entropy which is caused due to partitioning the training samples.

The gain on attribute (B) over training sample Sa is denoted as

$$\text{Gain (Sa, B)} \equiv \text{Entropy(Sa)} - \sum_{v \varepsilon Values(B)} \frac{|Sav|}{|Sa|} Entropy(Sav)$$

Where,

Values (B) are all possible values for Attribute B.
Sav is subset of sample set Sa for B has value v.
Entropy (Sa) is entropy for entire sample dataset Sa.
Entropy (Sav) denotes entropy for all subset Sav which is partition by attribute B.

The main drawback of ID3 is due to use of small sample dataset, over – classified and over – fitted result data is achieved. Single attribute is verified in iteration for decision making. It is also costly to perform classification on continues sample dataset because of getting the exact breakpoint for continues dataset it swamps many trees[13][14][15][16].

2.4 Naïve Bayesian Classifier

Navie bayes is probabilistic classifier use for classification based on Bayes rules. Cooper et. al. suggest that a Bayesian network have highest posterior probability which can be applied for classification of prior events . Consider an example of bayesian classifier to find the conditional probability C of each attribute A. Bayes rule is apply for calculating the conditional probability on instances of Attribute A1,…,An and predicting the highest posterior probability class.

It works depends on particular feature of the class instead of other features. Navie bayes classifier is faster for decision making. Bayes method involve learning hypothesis use for classification of every instances [17][18][19].

2.5 Random Forest

Breiman has proposed the Random Forest is ensemble learning method use for classification and regression. It create number of trees at training period with outputting the class. For each intermediate node in decision tree requires to take decision. Prediction of classification is done to get best class by merging the decision trees result.
The random forest becomes robust due to random input, random features and work with large feature space dataset [21][22][23].

2.6 Support Vector Machine

The SVM is statistical learning approach used in various applications such as gene analysis, patent classification , face recognition , and predicting longitudinal dispersion coefficients in natural rivers .
It is mainly use for classification and regression problems on small number of sample dataset and high dimension. SVM achieve higher performance with accuracy for classification by mapping the input data with feature space. This feature space linearly divided into various separated hyperplane[24][25][26][27][28][29][30].

2.7 C4.5

It is improvement over ID3 algorithm developed by Quinlan. It handle missing data and noisy data in better way as compare to ID3. Noisy data is created due to same attribute value in more than two examples. Due to error in data acquisition process and in data preprocessing step, attribute values becomes incorrect which produce wrong classification.

It also performs pre and post pruning of decision tree for avoiding over-fitting dataset [31][32].

3 Proposed Methodology

The methodology for structural identification of crystalline solids shall include:

i. Data collection

Data collection is important task for the desired research. Crystal structure data should include with all structural properties such as lattice parameters, Millar indices, diffraction intensities, electron densities, energies, packing fractions, phases present in the sample, atomic positions, bond lengths, coordination numbers, X – ray densities, porosities, and various framework type codes are essential for classification process.

ii. Feature extraction of crystal structure

One can also undertake the properties' study of different crystal structures. And perform feature extraction for further investigations.

iii. Classification of crystal structure.

With sufficient inputs one can apply classifier on preprocess sample dataset for getting optimal result. This optimal result will focus on classification accuracy as well as computational time accuracy.

iv. Identification of crystal structures

finally machine learning approach could be applied for identification of crystal structures.

4 Conclusion

This review monograph conclude that the classification techniques are useful in distinct areas. The machine Learning approaches are also suitable for structural identification purposes. The Machine Learning tools are having various best classifiers like Decision tree classifiers, SVM, C4.5 can be useful to work on prediction of structures. These Methods will improve the efficiency using appropriate attribute.

References

1. Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad, Reviews in Computational Chemistry, Volume 29, First Edition. © 2016 John Wiley & Sons, Inc.
2. Ethem Alpaydın. 2014 . Introduction to Machine Learning. The MIT Press, Third Edition.
3. Salvatore Ruggieri, Efficient C4.5, IEEE Trans. on Knlowedge and Data Engg., Vol 14, No.2, Mar 2002, pp. 438-444.
4. Ohbyung Kwon a, Jae Mun Sim ,Effects of data set features on the performances of classification algorithms Expert Systems with Applications ,40 (2013) 1847–1857.
5. Cristopher Carl Fischer. 2007 "A Machine Learning Approach to Crystal Structure Prediction", Doctoral Thesis. MIT press.
6. Ghanshyam Pilania , Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran and Ramamurthy Ramprasad, "Accelerating materials property predictions using machine learning", Sci. Rep. 3, 2013, 1-6.
7. D. Andrew Carr, Shujiang Yang, Mohammed Lach-hab, Iosif I. Vaisman, and Estela Blaisten-Barojas, " Machine Learning Approach for structure based zeolite classification", Microporous and Mesoporous Materials, 2008.

8. C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater., 5, 641 (2006). Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics.
9. G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater., 22, 3762 (2010). Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory.
10. Brodley, C. E., & Utgoff, P. E. Multivariate decision trees, Machine Learning, 19, 45-77.
11. Buntine, W., & Niblett, T. A further comparison of splitting rules for decision-tree induction. Machine Learning, 8, 75-86.
12. Quinlan J.R. 1986. Induction of Decision Trees. Machine Learning, 81-106.
13. Tom M. Mitchell, Machine Learning, MGH International, 1997, pp. 177-178
14. Zhongbo Zhang, Shuanghu Zhang , Simin Geng, Yunzhong Jiang, Hui Li, Dawei Zhang, Application of decision trees to the determination of the year-end level of a carryover storage reservoir based on the iterative dichotomizer 3, Electrical Power and Energy Systems 64 (2015) 375–383.
15. http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm
16. Badr Hssina, Abdelkarim Merbouha, ,Hanane Ezzikouri, Mohammed Erritali ," A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications ,2014, 13-19.
17. Hitesh et al., International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October -2013, pp. 955-963.
18. Ahmad Ashari, Iman Paryudi, A Min Tjoa, Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, International Journal of Advanced Computer Science and Applications,Vol. 4, No. 11, 2013, pp. 33-39]
19. http://cmasc.gmu.edu/publications_blaisten/90.pdf
20. Ghanshyam Pilania , Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran & Ramamurthy Ramprasad, Accelerating materials prop-

erty predictions using machine learning, www.nature.com\scientificreports ,sci. rep. 3 2810;DOI: 10.1038/SREP02810 (sept 2013),pp.1-6.

21. Detlef W.M. Hofmanna, Joannis Apostolakisb ,Crystal structure prediction by data mining, Journal of Molecular Structure 647 (2003) 17–39.
22. Dr. S.N. Kakarwal, P.D. Patni, A Review on Comparison of Iterative Dichotomiser 3 and Naïve Bayes Classifiers, IInd International conference on nano-structured materials and nano-composite, ICNM 2014
23. Facial expression recognition from image sequences using twofold random forest classifier, Xiaorong Pu, Ke Fan,Xiong Chen , Luping Ji , Zhihu Zhou [Neurocomputing] 168 (2015) 1173–1180.
24. V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York Inc., NY, USA, 1995.
25. I. Guyon, J. Weston, S. Barnhill, V.N. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1–3) (2002) 389–422.
26. C.-H. Wu, Y. Ken, T. Huang, Patent classification system using a new hybrid genetic algorithm support vector machine, Applied Soft Computing 10 (4) (2010) 1164–1177.
27. S. Chowdhury, J.K. Sing, D.K. Basu, M. Nasipuri, Face recognition by generalized two-dimensional FLD method and multi-class support vector machines, Applied Soft Computing 11 (7) (2011) 4282–4292.
28. Umi Kalsum Hassan, Nazri Mohd. Nawi, and Shahreen Kasim, Classify a Protein Domain using Sigmoid Support Vector Machine, IEEE 2014, International Conference on Information Science and Applications (ICISA), pp1-4.
29. Xiao Li Zhang a,b , Wei Chen c , BaoJian Wang , XueFeng Chen , Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization, Neurocomputing 167 (2015) 260–279.
30. Cristianini, N. and Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, 2000.
31. Carlos J. Mantas, Joaquín Abellán , Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data , Expert Systems with Applications 41 (2014) 4625–4637.
32. Quinlan J.R. 1993. C4.5: programs for machine learning. Morgan Kaufmann.