EXPRESS LETTER

Machine Learning to Instruct Single Crystal Growth by Flux Method*

Machine Learning to Instruct Single Crystal Growth by Flux Method *

Tang-Shi Yao(姚唐适)[1,2†], Cen-Yao Tang(唐岑瑶)[1,2†], Meng Yang(杨萌)[1,2†], Ke-Jia Zhu(朱恪嘉)[1,2], Da-Yu Yan(闫大禹)[1,2], Chang-Jiang Yi(伊长江)[1,2], Zi-Li Feng(冯子力)[1,2], He-Chang Lei(雷和畅)[4], Cheng-He Li(李承贺)[4], Le Wang(王乐)[1,2], Lei Wang(王磊)[1**], You-Guo Shi(石友国)[1,2**], Yu-Jie Sun(孙煜杰)[1,3,5**], Hong Ding(丁洪)[1,3,5]

[1]*Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing 100190*

[2]*University of Chinese Academy of Sciences, Beijing 100049*

[3]*CAS Centre for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing 100049*

[4]*Department of Physics and Beijing Key Laboratory of Opto-electronic Functional Materials and Micro-nano Devices, Renmin University, Beijing 100872*

[5]*Songshan Lake Materials Laboratory, Dongguan 523808*

(Received 30 April 2019)

Growth of high-quality single crystals is of great significance for research of condensed matter physics. The exploration of suitable growing conditions for single crystals is expensive and time-consuming, especially for ternary compounds because of the lack of ternary phase diagram. Here we use machine learning (ML) trained on our experimental data to predict and instruct the growth. Four kinds of ML methods, including support vector machine (SVM), decision tree, random forest and gradient boosting decision tree, are adopted. The SVM method is relatively stable and works well, with an accuracy of 81% in predicting experimental results. By comparison, the accuracy of laboratory reaches 36%. The decision tree model is also used to reveal which features will take critical roles in growing processes.

PACS: *81.10.−h*, *61.50.Ah*, *81.05.−t*, *89.20.Ff*   DOI: 10.1088/0256-307X/36/6/068101

Single crystals are vital prerequisite for extensive scientific research fields, such as condensed matter physics, surface science, lasers and nonlinear optics. Fundamental studies, e.g., on quantum Hall effect/fractional quantum Hall effect,[1] Wyle semimetal,[2−4] etc., all rely on the production of high-quality single crystals. Moreover, many useful experimental techniques are applied to single crystals exclusively.[5,6] Unfortunately it is not easy to grow single crystals owing to the complexity of interrelated factors such as temperature, composition ratios and atomic radius.[7−10] In particular, growth of ternary compounds is difficult due to the lack of ternary phase diagrams. Therefore, production of single crystals is notoriously labor-wasting and time-consuming. With the development of machine learning (ML) theory and faster calculation speed,[11−17] ML makes excellent forecast by learning from large databases and outperforms people in various fields.[18−23] Recently, ML tools have been widely used in material science and high-throughput computations.[9,24−33]

In this Letter, we focus on single crystal growth of ternary compounds by flux method, which is one of the most widely used methods in laboratories. We collect initial data from laboratory notebooks including growth temperature curve, composition elements, ratios, and flux. The quantity and the quality of the data used is of importance for generating successful ML models. Data quality issues that often need to be addressed include the presence of noise and outliers, missing, inconsistent or duplicate data, and data that are biased or, in some other way, unrepresentative of the phenomenon or population that the data are supposed to describe.[34] After excluding reactions with incomplete laboratory notebook entries, we obtain 775 complete experimental data from group-I, and 649 data with 272 different kinds remained after removing duplicate conditions. The single crystals in our data set of group-I are not only diverse, but also contain 65 elements. For further verification, we also obtain 163 data from group-Ⅱ, and 115 remained after removing duplicates, data size is shown in Fig. 1(a). The main purpose of our work is to help the laboratory to grow new materials. We concern about whether the model can predict samples of new varieties. In order to reflect the ability of the model in predicting new samples, the single crystal types among the training set, the validation set and the test set are different.

Using ML to judge whether a crystal can be successfully grown in given conditions is a problem of two-category classification. The labels of our model are whether the single crystal we need is successfully

*Supported by the National Key Research and Development Program of China under Grant Nos 2016YFA0401000 and 2017YFA0302901, the National Basic Research Program of China under Grant No 2015CB921000, the National Natural Science Foundation of China under Grant Nos 11574371, 11774399 and 11774398, the Beijing Natural Science Foundation (Z180008), and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No XDB28000000.

†Tang-Shi Yao, Cen-Yao Tang and Meng Yang contributed to this work equally.

**Corresponding authors. Email: wanglei@iphy.ac.cn, ygshi@iphy.ac.cn, yjsun@iphy.ac.cn

grown. The features origin from the growth conditions of the sample (e.g., raw material ratio, flux, maximum temperature, minimum temperature, cooling rate, maximum temperature residence time), composition of sample element physicochemical properties (such as elemental electronegativity, atomic radius, elements melting point, elemental volatility, position of the atom in the periodic table), sample properties (such as crystal space group, crystal mass), phase diagram extraction information (such as melting point at different ratios). The detailed features involved in our model are listed in Tables S1-S3 in the supplementary material. Some processing methods are taken to make the model more generalizable: evaluating the model with 10-fold crossing-validation and testing T-statistic hypothesis ensure the results to be reliable, the boosting method reduces the impact of uneven distribution of sample data point, principal component analysis (PCA) filters noise and reduces features. Considering the size of data, we adopted SVM, decision tree, random forest (RF) and gradient boosting decision tree (GBDT) to study this problem. These ML algorithms are suitable for dealing with classification problems where data size is small. The SVM algorithm maps the original feature space to a higher-dimensional Hilbert space through a kernel function to find the super plane of the classification.[35] Decision tree is an algorithm that continuously purifies the nodes of the decision tree by information theory to achieve the best classification.[36] RF[37] and GDBT[38] are the integrated algorithms of bagging and boost ideas applied to decision tree algorithms, respectively (for details see Text B in the supplementary material).

Figure 1(b) gives the scheme of our feedback mechanism as following steps: (i) Extract features from theories of crystal growth and experience. (ii) Train and test the model based on the selected features. (iii) Reassess features from model outcomes. (iv) Analyze the important factors which impact the process of crystal growth including nucleation and crystallization via the outcomes of decision tree. (v) Predict the results based on new experimental data, and add the new data with adjusted weights into the original dataset to retrain the model. Furthermore, new laboratory data can help assessing and improving our model. Accuracy rate, f1 score, recall rate and precision of successfully grown samples are four indicators used to evaluate the model.[34] The accuracy and the f1 score of the model represent the learning ability of the model. The higher the scores of these two indicators, the stronger the model analysis ability. A high recall rate for successful sample predictions indicates that it is not easy for a model to misjudge a successful condition as a failure condition. The accuracy of predictions for successful samples represents the success rate of growth according to the conditions provided by the model. In order to better apply the model to crystal growth and study the characteristics of crystal growth, we compared outcomes of different models. Outcomes of SVM, decision tree, RF and GBDT on group-I and group-Ⅱ test sets are shown in Figs. 1(c) and 1(d), respectively. Average statistical results of 10-fold cross-validation of SVM, decision tree, RF and GDBT on group-I and group-Ⅱ datasets are shown in Figs. 1(e) and 1(f). The SVM model has stable and better performance. Therefore, we choose SVM to assist in growing crystals in experiments. And decision tree is used to clarify feature importance, which is more visualized. Group I studied the single crystal growth of ternary compounds. Since the data amount of the ternary compounds of group Ⅱ is relatively small, group-Ⅱ data contain binary compounds. The SVM trained model on data of groups I and Ⅱ has accuracies of 81% and 78%, respectively.

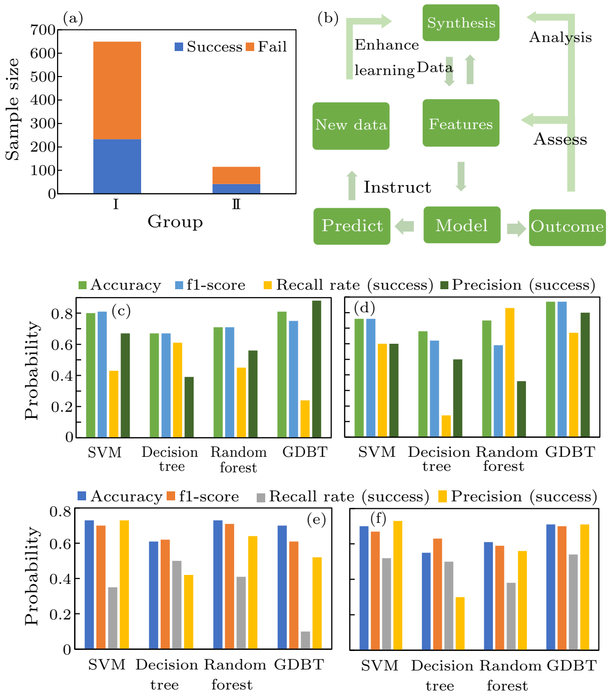

Fig. 1. Illustration of datasets and models. (a) Data size from group I and group Ⅱ. (b) Scheme of the feedback mechanism. (c)–(d) Outcomes of SVM, decision tree, random forest and GDBT on group-I and group-Ⅱ test sets, respectively. The text in the boast represents the label category. (e)–(f) Average statistical results of 10-fold cross validation of SVM, decision tree, random forest and GDBT on group-I and group-Ⅱ datasets, respectively. The text in the boast represents the label category.

Table 1. Outcomes of the SVM model.

Group I test data				
	Accuracy	Recall rate	F1-score	Data size
Failure	0.84	0.93	0.88	72
Success	0.67	0.43	0.53	23
Average/sum	0.80	0.81	0.80	95
Group Ⅱ test data				
Failure	0.83	0.83	0.83	23
Success	0.60	0.60	0.60	10
Average/sum	0.76	0.76	0.76	33

To demonstrate the performance of our model, confusion matrixes are shown in Table 1. Table 1 shows the SVM result on group-I test set, reaching average accuracy, average recall rate, f1-score of 81%, 81%, and 80%. The corresponding values on group Ⅱ are 76%, 76%, and 76%, respectively, as also shown in Table 1. The stability of our model is strong, and the prediction results of different laboratories are relatively consistent. It is worth mentioning that the test set contains totally different samples from training set, and thus revealing great generalization capability of our model. From all above, it is strongly believed that our SVM model is pretty promising in instructing single crystal growth condition exploration. What we have already carried out is predicting outcomes for given growth conditions. The precision on failed samples of our model is high, so discard the conditions which the model considers as failure. New data are taken from the laboratory as extra data. In these data, the success rate is 34%. Meanwhile, if we only choose the conditions recommended by our model, the success rate reaches as high as 71%. The accuracy of ML's judgment on success conditions is significantly higher than that of human, which largely avoids time and money wasting.

A successful model should both increase synthesis success rate and give physical insight. There is no intersection of features and convolution in the decision tree training process, so the role of independent features in the whole learning process is visible. For this reason, we use decision tree with 32 features (shown in Table 2) to train on data from both the groups and present the feature importance in Fig. 2. For convenience of discussion, we define the compound we studied as $A_xB_yC_z$ and the flux used as M, where A, B, and C are elements, x, y and z are their relative ratios. Usually A is an alkali metal, alkaline earth metal or rare earth element, B is a transition metal, and C belongs to 3rd–6th main groups. As shown in Fig. 2(a), the difference of electronegativity between flux and composition elements plays the most important role. Meanwhile, the temperature curve during the synthesis process is considered to be important, without doubt. And those factors illustrate elements themselves, such as composition ratio, element radius and dissolve point in flux, strongly influence the synthesis procedure. However, density, the usually ignored factor by experimentalists, has a significant impact on single crystal growth. From group-Ⅱ dataset, the similar result is presented.

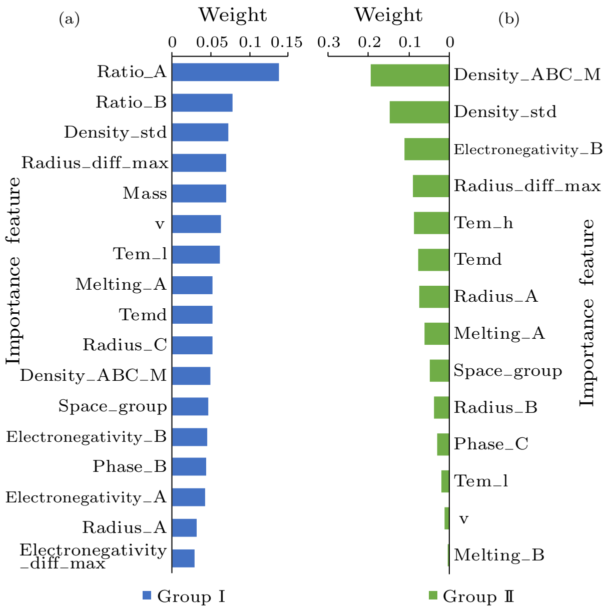

Fig. 2. Feature importance from the decision tree model: (a) feature importance of group-I data, (b) feature importance of group-Ⅱ data.

Table 2. Feature description, i.e., the meaning of the features from decision tree model in Fig. 2.

Feature name	Description
Phase_A/B/C	The lowest temperature at which A/B/C is completely dissolved
Phase_max	The maximum of phase_A, phase_B and phase_C
Boiling_M	Boiling point of flux M
Melting_M	Melting point of flux M
Electronegativity_M	Atomic electronegativity of flux M
Radius_M	Atomic radius of flux M
Radius_A/B/C	Atomic radius of A/B/C
Electronegativity_A/B/C	Atomic electronegativity of A/B/C
Melting_A/B/C	Melting point of A/B/C
Tem_h	Maximum temperature
Ratio_A/B/M/N	Molar ratio of raw material A/B/M/N and C
Mass	Molecular mass of the sample
Space_Group	Space group of samples
Density_std	Standard deviation of all elemental densities
Temd	Cooling rate
v	Cooling rate
Tem_l	centrifugal temperature
Radius_diff_max/ave	Maximum/mean of radius difference between elements and flux M
Electronegativity_diff_max	Maximum of electronegativity difference between elements and flux M

The tree of single crystal growth process from group I is shown in Fig. 3. The conditions collected are artificially selected for growing single crystals, that is to say, these growth conditions are considered to be successful by human experience. In the absence of ternary phase diagram, experimentalists generally use a binary phase diagram as an alternative to select growth conditions. Therefore, the tree graph can show some factors that are often overlooked or less considered in laboratories. Several rules are summarized from the decision tree. Some of them are well understood but re-discovered by ML, for example, better single crystal growth is associated with (a) lower cooling rate and (b) lower melting point of B in flux. ML also discovers some rules without clear explanatory theory, which are of particular interest. For example, ML suggests that it is difficult to grow single crystals: (a) the x/z and y/z values are relatively large, (b) the average difference between the liquid phase density of flux and the components of single crystal (A, B, and C defined above) is large. Furthermore, it is recommended to choose a flux with a smaller maximum of absolute electronegativity differences among A, B and C.

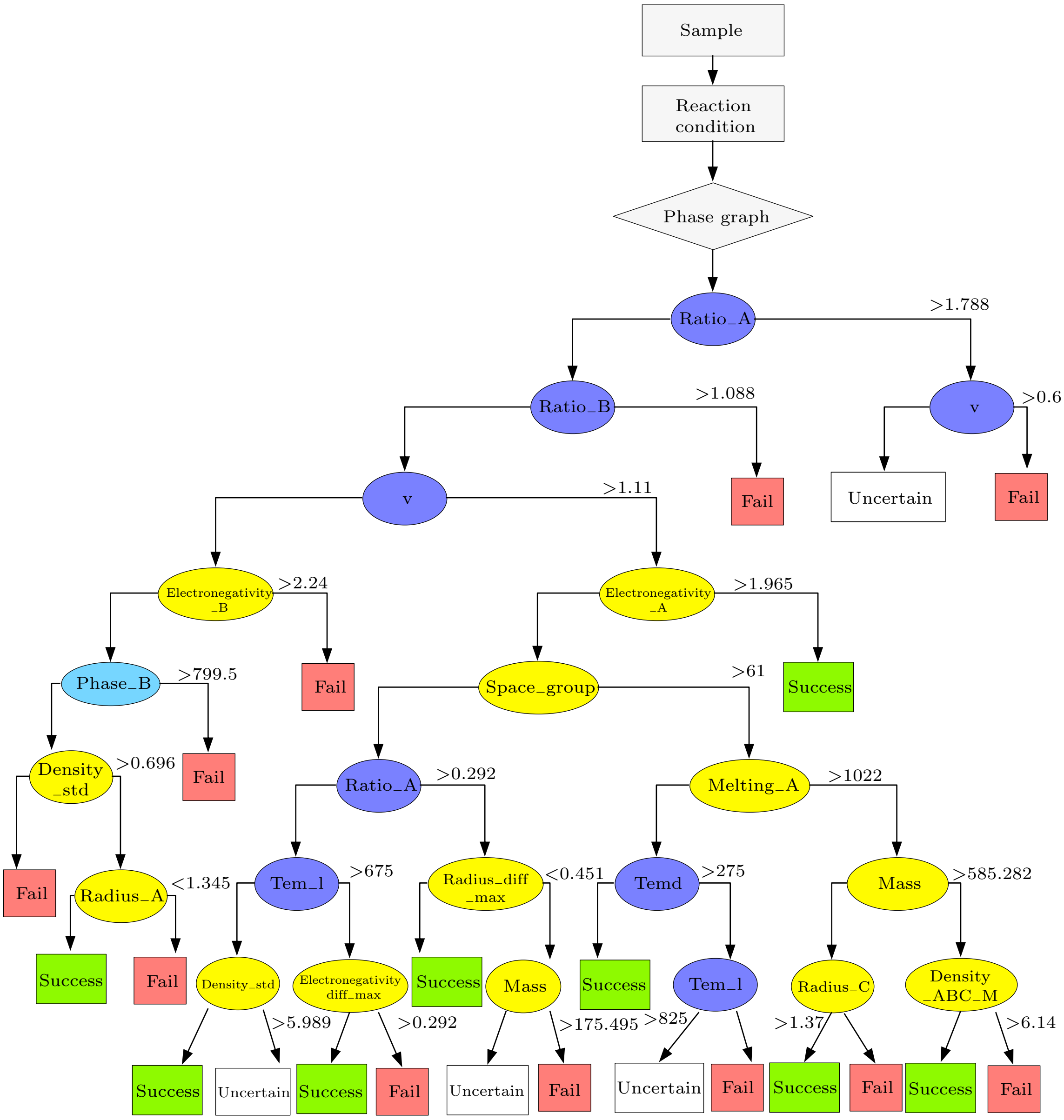

Fig. 3. Tree of single crystal growth process. The decision tree of group I. the yellow feature is the physicochemical properties of the grown sample and its constituent elements, the blue features are extracted from the binary phase diagram, the purple features are derived from the experimental records, which is the growth condition of the sample, and the green color is represented in this branch, good sample is grown, and red indicates that no good sample is obtained.

In summary, we have applied the SVM model to train and test laboratory single crystal growth data. The success rate of our model is twice the rate of manual conditions. Meanwhile, the decision tree model is used to assess the importance of features and provides some new rules about single crystal growth without clear explanation. Although the samples we used to test our models are limited, based on the scientific and rigorous statistical evaluation methods and diversity samples, we have an enough assessment of the effects of the models (for details, see Text B in the supplementary material). In other words, our model can be applied to other materials with university and generality. The ML is proved to be helpful for instructing single crystal growth, despite the limited size of the data base and method of growth. It is easy to extend our study to other growth methods if corresponding data are available. With continuous growing of the databases, the ML will be better at the prediction of single crystal growth and uncover more essential rules in the process.

References

[1] Klitzing K v, Dorda G and Pepper M J P R L 1980 *Phys. Rev. Lett.* **45** 494

[2] Xu S Y, Belopolski I, Alidoust N, Neupane M, Bian G, Zhang C, Sankar R, Chang G, Yuan Z and Lee C C 2015 *Science* **349** 613

[3] Weng H, Fang C, Fang Z, Bernevig B A and Dai X 2015 *Phys. Rev. X* **5** 011029

[4] Chen X 2015 *Sci. Chin. Mater.* **58** 675

[5] Binnig G and Rohrer H 1983 *Surf. Sci.* **126** 236

[6] Damascelli A 2004 *Phys. Scr.* **2004** 61

[7] Fisher I, Shapiro M and Analytis J 2012 *Philos. Mag.* **92** 2401

[8] Canfield P C and Fisk Z 1992 *Philos. Mag. B* **65** 1117

[9] Raccuglia P, Elbert K C, Adler P D, Falk C, Wenny M B, Mollo A, Zeller M, Friedler S A, Schrier J and Norquist A J 2016 *Nature* **533** 73

[10] Nielsen J W and Dearborn E F 1958 *J. Phys. Chem. Solids* **5** 202

[11] Kohavi R 1996 *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon 1996) pp 202–207

[12] Shevade S K, Keerthi S S, Bhattacharyya C and Murthy K R K 2000 *IEEE Trans. Neural Netw.* **11** 1188

[13] Shalev-Shwartz S, Singer Y, Srebro N and Cotter A 2011 *Math. Programming* **127** 3

[14] Cherkassky V and Ma Y 2004 *Neural Networks* **17** 113

[15] Joachims T 1998 *Technical Report* SFB 475 (Komplexitätsreduktion in Multivariaten)

[16] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K and Kuksa P 2011 *J. Mach. Learn. Res.* **12** 2493

[17] Ren S, He K, Girshick R and Sun J 2015 *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems* (Montreal, Canada 7–12 December 2015) vol 1 p 91

[18] Churchland P S, Sejnowski T J and Poggio T A 1992 *The Computational Brain* (Cambridge: MIT Press) p 544

[19] Magerman D M, 1995 *Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics* pp 276–283

[20] Leslie C, Eskin E and Noble W S 2002 *Pacific Symposium on Biocomputing* **7** 564

[21] Zhang H, Berg A C and Maire M 2006 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (New York, USA 17–22 June 2006)

[22] Brill E 1995 *Comput. Linguistics* **21** 543

[23] Hoo-Chang S, Roth H R, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D and Summers R M 2016 *IEEE Trans. Med. Imaging* **35** 1285

[24] Zhang Y and Ling C 2018 *npj Comput. Mater.* **4** 25

[25] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K R 2017 *Nat. Commun.* **8** 872

[26] Snyder J C, Rupp M, Hansen K, Müller K R and Burke K 2012 *Phys. Rev. Lett.* **108** 253002

[27] Ren F, Ward L, Williams T, Laws K J, Wolverton C, Hattrick-Simpers J and Mehta A 2018 *Sci. Adv.* **4** eaaq1566

[28] Pillong M, Marx C, Piechon P, Wicker J G, Cooper R I and Wagner T 2017 *CrystEngComm* **19** 3737

[29] Wicker J G and Cooper R I 2015 *CrystEngComm* **17** 1927

[30] Zhou Q, Tang P, Liu S, Pan J, Yan Q and Zhang S C 2018 *Proc. Natl. Acad. Sci. USA* **115** E6411

[31] Curtarolo S, Hart G L, Nardelli M B, Mingo N and Sanvito S and Levy O 2013 *Nat. Mater.* **12** 191

[32] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 *Nature* **559** 547

[33] Natarajan A R and Van der Ven A 2018 *npj Comput. Mater.* **4** 56

[34] Tan P N, Steinbach M and Kumar V 2006 *Introduction to Data Mining* (New York: Pearson) vol 1 chap 2 p 41

[35] Vapnik V N 1999 *IEEE Trans. Neural Netw.* **10** 988

[36] Quinlan J R 1986 *IEEE Int. Workshop Mach. Learn. Signal Process.* **1** 81

[37] Ho T K 1995 *International Conference on Document Analysis and Recognition* (Montreal, Quebec, Canada 14–16 August 1995)

[38] Friedman J H 2001 *Ann. Stat.* **29** 1189