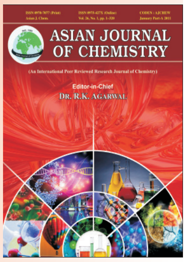
Asian Journal of Chemistry; Vol. 26, No. 17 (2014), 5839-5841

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2014.18217



Novel Framework for Improving the Desired Structure Prediction on Imbalance Data Set†

Hui Sun[1], Qingji Guan[2], Qiaohong Hao[2], Jun Kong[2], Yinghua Lu[1,*] and Miao Qi[2,*]

[1]College of Humanities & Sciences of Northeast Normal University, Changchun, P.R. China
[2]School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, P.R. China

*Corresponding author: Fax: +86 431 84536331; Tel: +86 18686501175; E-mail: qim801@nenu.edu.cn

AJC-15805

Desired target prediction has important significance in rational synthesis of materials. In this paper, taken (6,12)-ring-containing structure as the desired target, a novel framework is proposed for improving the prediction performance on the microporous aluminophosphates (AlPOs) database. In contrast to existing data processing techniques for class imbalance problem, the proposed framework first explores the intra-class distribution of majority class by clustering such that multiple specific models can be established according to the distribution. The main advantage is that one model can characterize the clustered data well. Then, Q times n-fold cross-validation procedure is applied to assess the prediction performance. Finally, we compare the proposed framework with two existing data processing procedures. The comparison results demonstrate that the desired target predictions can achieve improved performance remarkably.

Keywords: Aluminophosphates synthesis, Prediction, Affinity propagation clustering, Class imbalance.

INTRODUCTION

Nowadays, zeolites and related microporous materials have been highly used in the petroleum industry due to their excellent specificities[1]. A better understanding of the relationships between the synthetic factors and the resulting structures is very important for rationalizing the synthesis of the target zeolitic materials. However, the synthetic processes of these materials are very complex and governed by a number of factors. Therefore, it is a challenge task to exploit the relationships.

Data mining and machine learning technologies have been widely used to speed up the discovery and modeling the relationship[2,3]. More recently, many research works have been reported on aluminophosphate synthesis database[4-8]. These works have contributed to the understanding of the synthesis mechanism. Li *et al.*[4] used support vector machine (SVM) to predict (6,12)-ring-containing microporous aluminophosphates and gave the best combination of synthetic factors[4]. Partial least squares and logistic discrimination (PLS-LD) methods were presented to predict the generation of microporous aluminophosphate aluminophosphate$_4$-5[5]. Besides, four re-sampling methods were proposed to deal with the problem of class imbalance. For the sake of examining the significant synthetic factors affecting the formation of (6,12)-ring-containing structure, an integrated feature selection method was proposed based on random subspace method in our previous work[8].

In this paper, a novel framework is proposed to predict the desired target on imbalance data set of aluminophosphates database. Different to most existing predictive models as well as the above mentioned methods, intra-class distribution of the majority class is first investigated adequately by affinity propagation (AP) clustering. Then, based on the clustering results, Q times n-fold cross-validation (CV) is applied to assess the predictive performance. More specifically, several classical classifiers are used to measure the validity of proposed framework. The results are judged on the numerical prediction of (6,12)-ring-containing structure.

EXPERIMENTAL

The experimental data set is from aluminophosphate database and (6,12)-rings structure is used as predicted target[9]. In particular, twenty one synthetic factors and 1279 items are selected as experimental data set. In this data set, 398 samples that can produce the (6,12)-ring-containing structure are called positive samples. The other 881 samples are the negative samples.

Method: Affinity propagation clustering algorithm has gained increasingly popularity in recent years as an efficient and fast clustering algorithm[10]. Compared with classical k-means, k-centres and Fuzzy C-Means clustering methods, affinity propagation algorithm needs not to predefine the number of clusters and to initial the exemplars and has fast

†Presented at 2014 Global Conference on Polymer and Composite Materials (PCM2014) held on 27-29 May 2014, Ningbo, P.R. China

computation. Given a data set, affinity propagation algorithm can be described as:

```
Algorithm (affinity propagation clustering):
Begin initialize $t_{max}$, $t \leftarrow 0$, $a(i, k) \leftarrow 0$, $i, k = 1, ... n$.
Compute the similarity matrix s
        $s(i,k) \leftarrow -\|x_i - x_k\|^2$
do
        $t \leftarrow t+1$
    $r(i,k) \leftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i,k') + s(i,k')\}$
    $a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \text{ s.t.} i' \notin \{i,k\}} \max\{0, r(i',k)\}\}$
Until $t = t_{max}$ or convergence criterion met. Return $\{x_k\}$,
satisfying $r(k,k) + a(k,k) > 0$.
```

Performance measure: The performance of a prediction algorithm is assessed by F-measure[5], which is an integrated measure and considers both two classes accuracy simultaneously. It is a very effective metric for measuring imbalance data set.

RESULTS AND DISCUSSION

We observed that the data set is obviously unbalanced. To better handle this problem, this paper proposes a novel framework to improve the predictive results, which considers the data distribution of the majority class. It can complete the predictive task by establishing multiple classifiers such that each classifier has its particular discriminant ability. The framework can be depicted as: (1) Cluster majority class into C subsets by affinity propagation clustering algorithm. (2) Classify the each subset with the minority class using Q times n-fold CV. (3) Average the C results as the final predictive results.

Parameters setting: The negative samples are divided four subsets by using affinity propagation algorithm. Accordingly, C is set to 4. Considering the sizes of subsets and positive samples, we set Q to 10 and n = 3, 5, 7, 10, respectively. Moreover, SVM with Gaussian kernel, Adaboost, K-Nearest Neighbor algorithm (KNN), Back Propagation network (BP), Classification and Regression Tree (Cart), Iterative Dichotomizer-3 (ID3) and PLS-LD[11] are adopted to complete the prediction task. For KNN algorithm, the predictive result will be affected by neighbor size k. The results of different k are shown in Table-1. We can clearly see that they reach best performances when k = 1. Therefore, the value of k is set to 1 in the following experiments.

TABLE-1
DEPENDENCE OF PREDICTIVE RESULTS ON THE NEIGHBOR SIZE k (%)

CV	k = 1	k = 3	k = 5	k = 7
3-fold	64.11	62.18	59.59	57.52
5-fold	43.44	33.57	28.75	23.85
7-fold	29.55	22.48	19.40	16.18
10-fold	18.52	15.13	13.15	12.25

Similar, for PLS regression in PLS-LD, the number of PLS components is determined by testing different components (d) in terms of F-measure. The changed trend with different d is depicted in Fig. 1. Obviously, all curves take on ascendant trends when increasing d from 1 to 6. The n-fold curves reach the top with 93.28, 93.25, 93.37 and 93.42 when d = 12, 14, 14, 12, respectively. When d = 12, the values of 5-fold CV and 7-fold CV are 93.04 and 93.35 %, which are lower the top points by merely 0.21 and 0.02 %, respectively. Likewise, all curves show the downward trends when d > 14. Consequently, the number of PLS components is set as d = 12.

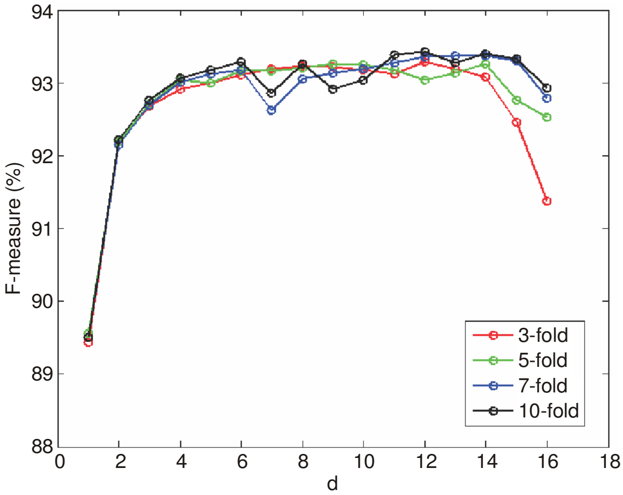


Fig. 1. Predictive performances with different d

After setting the optimal parameters C, k and d, the predictive performances of different classifiers are evaluated in terms of F-measure. The comparison results are shown in Fig. 2. It is clear that the performance of KNN descends rapidly with increasing n. It shows the worst results and gets the highest F-measure value only 64.11 %. However, the performances of other classifiers are affected by n rarely. SVM, Adaboost and PLS-LD give the best results compared with others. For 10-fold CV, SVM, Adaboost and PLS-LD exhibit the best performances with 94.50, 92.12 and 93.42 %, respectively. BP, Cart and ID3 give medium performances. Their F-measure values are 87.91, 87.80 and 82.15 %, respectively, which is lower than the highest result by 6.59, 6.70 and 12.35 %, respectively. The good predictive results indicate that the proposed framework based on clustering is feasible and can reach satisfied predictive results.

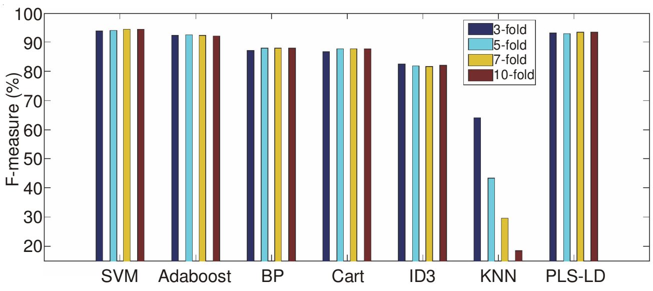


Fig. 2. Performance comparison of different classifiers

Comparison results: The main contribution of this study is that a novel framework for prediction is to propose rather than only select the good classifier. For further demonstrating the superiority of the proposed framework, we compare it with two general data processing modes: the first one does not consider the class imbalance problem and adopts n-fold CV directly (called Mode 1)[12]. The second is to select randomly the same number of samples with the minority class from the negative samples to complete the prediction task (called Mode 2)[4,8]. Accordingly, the proposed framework is called as Mode 3. The F-measure comparisons of Mode 1 and Mode 2 with different classifiers are first shown in Figs. 3 and 4. Similarly, SVM, Adaboost and PLS-LD keep the best performances for both modes. However, all results are lower than 90 %. Next, we compare the F-measure of the three modes. The visual comparison results of different modes for 7-fold CV are illustrated in Fig. 5. Obviously, Mode 3 exhibits the best predictive performance for each classifier, followed by Mode 2. Mode 1 gives the worst results. Taken Adaboost as an example, the predictive result of Mode 3 is 92.33 % and higher than Mode 1 and Mode 2 by 19.46 and 10.95 %, respectively.

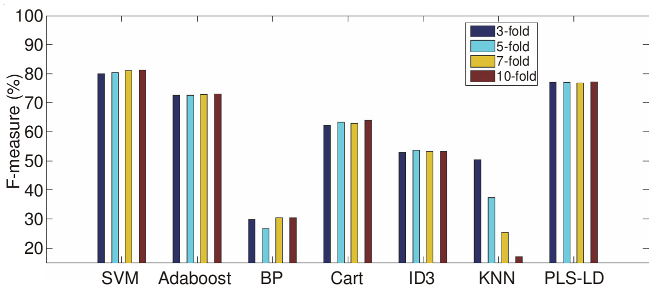


Fig. 3. Performance comparison of different classifiers for mode 1

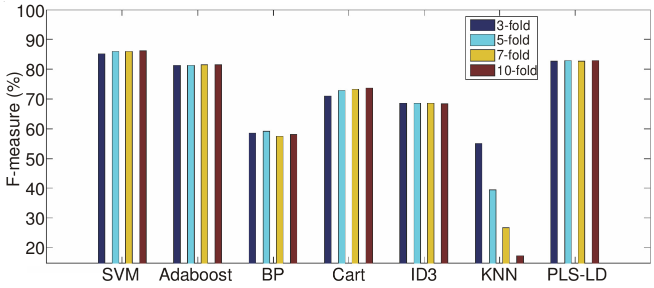


Fig. 4. Performance comparison of different classifiers for mode 2

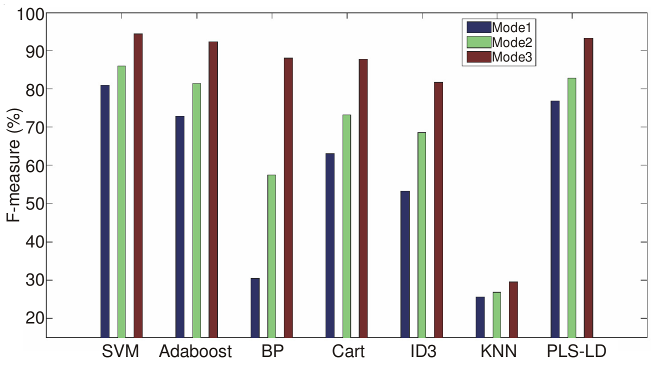


Fig. 5. Comparison of different methods

Analyzing the reason of good performance in Mode 3 is that it considers the data internal structure. By affinity propagation clustering, the samples with similarity are clustered in each subset. It can reduce the training burden of classifier with similar samples and make better classification accuracy. In other word, the trained classifier for each subset is specific. For Mode 1, it does not consider class imbalance. The established model may be apt for the majority class and results in the bad predictive results for the minority class. Mode 2 adopts multiple random sampling to deal with the imbalance problem. However, the samples in each selection might be dispersed such that the classifier can not model the training samples accurately.

Conclusion

A novel framework has been proposed for predicting the desired structure. Different from general data processing modes, the data structure is analyzed by clustering when the class imbalance problem occurs. Compared with two existing modes, the proposed framework achieves the best predictive results, which can provide a significant reference for aluminophosphates rational synthesis.

REFERENCES

1. H. Lee, S.I. Zones and M.E. Davis, *Nature*, **425**, 385 (2003).
2. X.D. Liu, Y.H. Xu, J.H. Yu, Y. Li, W. Zeng, C. Chen, J.Y. Li, W.Q. Pang, R.R. Xu and Y. Xu, *Chem. J. Chinese Univ.*, **24**, 949 (2003).
3. L.A. Baumes, M. Moliner and A. Corma, *QSAR Comb. Sci.*, **26**, 255 (2007).
4. J.Y. Li, M. Qi, J. Kong, J.Z. Wang, Y. Yan, W.F. Huo, J.H. Yu, R.R. Xu and Y. Xu, *Micropor. Mesopor. Mater.*, **129**, 251 (2010).
5. M. Qi, Y.H. Lu, J.Z. Wang and J. Kong, *Mol. Inform.*, **29**, 203 (2010).
6. W.F. Huo, N. Gao, Y. Yan, J.Y. Li, J.H. Yu and R.R. Xu, *Acta Phys. Chim. Sin.*, **27**, 2111 (2011).
7. J.S. Li, Y.H. Lu, J. Kong, N. Gao, J.H. Yu, R.R. Xu, J.Z. Wang, M. Qi and J.Y. Li, *Micropor. Mesopor. Mater.*, **173**, 197 (2013).
8. M. Qi, J.S. Li, J.Z. Wang, Y.H. Lu and J. Kong, *Ind. Eng. Chem. Res.*, **51**, 16734 (2012).
9. http://zeobank.jlu.edu.cn.
10. B.J. Frey and D. Dueck, *Science*, **315**, 972 (2007).
11. D.V. Nguyen and D.M. Rocke, *Bioinformatics*, **18**, 39 (2002).
12. J. Manuel Serra, L. Allen Baumes, M. Moliner, P. Serna and A. Corma, *Comb. Chem. High Throughput Screen.*, **10**, 13 (2007).