molecular informatics

DOI: 10.1002/minf.200900072

Novel Approach for Clustering Zeolite Crystal Structures

M. Lach-hab,[a] S. Yang,[a] I. I. Vaisman,[a, b] and E. Blaisten-Barojas*[a, c]

Abstract: Informatics approaches play an increasingly important role in the design of new materials. In this work we apply unsupervised statistical learning for identifying four framework-type attractors of zeolite crystals in which several of the zeolite framework types are grouped together. Zeolites belonging to these super-classes manifest important topological, chemical and physical similarities. The zeolites form clusters located around four core framework types: LTA, FAU, MFI and the combination of EDI, HEU, LTL and LAU. Clustering is performed in a 9-dimensional space of attributes that reflect topological, chemical and physical properties for each individual zeolite crystalline structure. The implemented machine learning approach relies on hierarchical top-down clustering approach and the expectation maximization method. The model is trained and tested on ten partially independent data sets from the FIZ/NIST Inorganic Crystal Structure Database

Keywords: Zeolites · Framework type · Clustering · Crystal structure · Unsupervised learning · Computational chemistry

1 Introduction

Data clustering and other unsupervised machine learning algorithms are widely used in a variety of disciplines, including molecular informatics, bioinformatics, spectroscopy, astrophysics and materials science. Applications such as pattern recognition, data analysis in commerce, image processing have recently seen important advances when employing data clustering algorithms. Clustering algorithms are unsupervised learning classification methods that search for similarities between entries in a data set and segments data collections into subsets displaying similarities.[1,2] Therefore, clustering is the process of separating data subsets from each other without knowing a priori the number of classes into which the existing data are to be distributed. This process contrasts with classification algorithms, which are supervised learning methods that rely on a predefined number of well established classes. Although there are a wide variety of rules for assessing the degree of dissimilarity between elements of a data set assigned to the respective clusters, it is instructive to think about Euclidian square distances between these elements as a measure of dissimilarity. Most commonly, distances are calculated in a multidimensional space spanned by the properties (attributes or features) assigned to each data entry.

Two important ways of performing clustering are hierarchical and partitional. Hierarchical clustering proceeds successively by either splitting large clusters into two smaller ones (top-down) or merging small clusters into larger ones (bottom-up). With this procedure, a hierarchy of nested clusters is generated and commonly represented by a rooted binary tree that is graphically displayed with a dendrogram. Partitional clustering divides the data set into a number of disjoint clusters containing instances with strong similarities while the instances in different clusters are highly dissimilar. A commonly used partitional clustering method is the Expectation Maximization (EM),[3] which iterates two basic steps. In the expectation step a soft assignment of the cluster centroid is performed based on the similarity matrix and in the maximization step the similarity matrix elements are weighted for maximum dissimilarity to update the estimates. One of the important advantages of the EM algorithm over other clustering methods such as PCA,[1,4] co-training,[5,6] *k*-means,[7–9] spectral clustering[1] among others is its ability for determining the optimal number of clusters.

In this work we use a hybrid hierarchical-partitional clustering approach for identifying clusters of framework types (FT) of zeolite crystalline structures. Each cluster represents a "super-framework" that combines similar FTs. The clustering procedure relies on our recently developed set of features that includes topological, physical and chemical characteristics of different framework types.[10] This feature set

[a] *M. Lach-hab, S. Yang, I. I. Vaisman, E. Blaisten-Barojas*
Computational Materials Science Center
George Mason University
MSN 6A2, Fairfax, VA 22030, USA
phone: 011-703-9931988
fax /011-703-9939330
**e-mail: blaisten@gmu.edu*

[b] *I. I. Vaisman*
Department of Bioinformatics and Computational Biology
George Mason University
MSN 5B3, Manassas, VA 20110, USA

[c] *E. Blaisten-Barojas*
Department of Computational and Data Sciences, George Mason University
MSN 6A2, Fairfax, VA 22030, USA

was generated for supervised classification yielding accuracies of up to 99%. Now we employ the same feature vector and EM to group together thirteen FTs into four groups of FTs and denominate each group as a FT attractor. Data used to generate the 9-dimensional feature vector are part of the FIZ/NIST Inorganic Crystal Structure Database (ICSD).[11] This paper is organized as follows. Section 2 describes the features, the zeolite data and the clustering methodology. Section 3 gives a detailed discussion of the results and Section 4 concludes this paper.

2 The Feature Vector

Zeolites are microporous aluminosilicate crystals with a structurally supporting network of TO4 units (T is a tetrahedrally coordinated atom to four oxygens) filled with exchangeable cations and the adsorbent phase (mostly water). Zeolites are widely used for adsorption, ion exchange, heterogeneous catalysis, and in a number of emerging areas such as biomedical technology, sensors, and solar energy conversion. Zeolites, with the diversity of their natural forms, are among the most abundant mineral species. In addition to about forty species occurring naturally, hundreds of other zeolites have been synthesized. According to the Structure Commission of the International Zeolite Association (IZA), there are 191 distinct networks.[12] The distinction is based on as many perfect, theoretically prescribed framework types for identifying the network of TO4 building units. Each network pattern replicates periodically giving rise to well-organized arrays of channels that comprise topological characteristics specific to different zeolites. Thus, the FT of a zeolite structure is a topological signature that identifies the network connectivity of the TO4 building units. Each FT is designated by a unique three-capital-letter code.

The ICSD contains structural information assembled from publications, mostly based on X-ray data and published in the literature. For this work we have used features corresponding to 1034 zeolite crystalline structures from this database.[13] The selection of the zeolite dataset size is based on pre-processing data entries in the ICSD and kipping those possessing framework types with population equal or larger than 19.[14] In fact, zeolites not included in this work populate framework types with 14 or less entries, which is a small number for the machine learning analysis envisioned in this article. These 1034 zeolite entries will be referred in the following paragraphs as "instances." These instances are distributed unevenly between thirteen FTs (CHA, EDI, FAU, FER, HEU, KFI, LAU, LTA, LTL, MFI, MOR, NAT, RHO) as shown in Figure 1.

The purpose of this work is to discover major trends that gather FTs into super-frameworks grouping around a number of native FTs. In order to accomplish this goal, a set of nine properties (9-D feature vector) is employed in our clustering model of zeolites. We use descriptors of

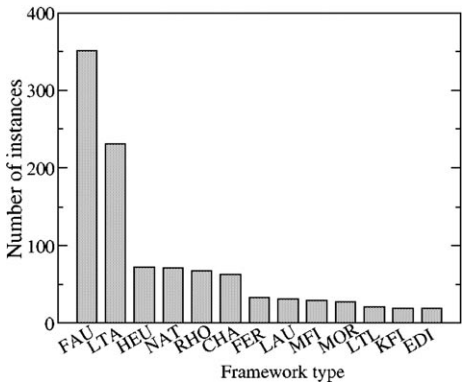

Figure 1. Number of instances in each of the framework types considered in the clustering model.

topological, physical and chemical nature that were originally constructed for supervised classification models.[10,14] Four topological descriptors are calculated using the Delaunay tessellation statistical geometry approach where nearest-neighbor points in 3-D space are connected by edges of Delaunay simplices. These descriptors include the tetrahedrality index T and the volume of largest inscribed sphere V.[15] The corresponding features are averages over these quantities generated for each zeolite entry in the ICSD pertaining to both first ($<T_1>$, $<V_1>$) and second tessellation shells ($<T_2>$, $<V_2>$).[13] We note that each instance displays tens of thousands of Delaunay simplices. The remaining five physical and chemical descriptors are: framework density, *FD* defined as number of T-atoms per 1000 Å^3, concentration of Si, [*Si*], concentration of Al, [*Al*], volume of the normalized reduced cell of the crystal, V_c, and its skewness, *s* defined as the deviation of the mean of the lattice angles from 90°.

Analysis of the data is done on standardized features, which are normalized as

$$x_i = (y_i - y_{\min})/(y_{\max} - y_{\min}), \quad (1)$$

where y_i are the values of the i^{th} nonnormalized feature and $y_{\min}$, $y_{\max}$ are the minimum and maximum values of that feature in the data set of 1034 instances.

3 Clustering into Framework-Type Attractors

The clustering method adopted is a top-down divisive approach starting from ten different balanced data sets in which each of the thirteen zeolite FTs is equally represented by 19 instances. Therefore, each of the ten initial data sets contains 247 instances and is generated by selecting at random 19 instances for each of the thirteen FTs. Dividing each set into two clusters, and then dividing again each of the two clusters into two additional clusters builds a cluster hierarchy. After two divisive steps the clustering procedure is terminated. Each of the two division processes is performed using the EM algorithm. The dissimilarity mea-

sure in EM is the matrix of squared Euclidian distances between instances x_i and x_j in a 9-D feature space:

$$d(x_i, x_j) = \sum_{\alpha=1}^{9} (x_{i,\alpha} - x_{j,\alpha})^2 \quad (2)$$

The first step of the hierarchy starts with unsupervised clustering by EM random selection of two cluster centroids from the data. Once these two clusters are separated with EM as implemented in the Weka package,[2,16] then two routes were adopted. In one, nothing is done and the route is referred to as unsupervised. The second route, referred to as semisupervised, adds complexity to the problem by considering *labels*; each label is a FT that does not participate as a feature in the clustering algorithm. This is achieved by adding a constrain to the first step of the clustering hierarchy in which the 19 instances that carry the same label should be grouped together into one of the two clusters. The process is equivalent to a semisupervised mechanism that leads to two clusters containing CHA, LTA, FAU, KFI, RHO in one cluster and EDI, FER. HEU, LAU, LTL, MFI, MOR and NAT in the other. As a reference, the two clusters without constrain continue to be monitored, and such unsupervised method is compared to the semisupervised results.

In the second step of the clustering procedure each of the previous clusters is divided into two with the EM algorithm without constrains. The final four clusters are shown in Table 1 for the unsupervised and semisupervised approaches. Results in the table are mean and standard deviation for ten datasets. *Label specificity* is defined as the percentage of instances with one label that clustered in each of the final four clusters. As seen from the table, the unsupervised method yields ten FTs with specificity one or about one, while the semisupervised method improves the almost perfect grouping of labels to thirteen. These four clusters are FT attractors, which we denominate L, F, M and H. Here, each letter corresponds to the FT with highest representation. As shown in Table 1, the L-attractor has LTA as the native component. The F-attractor corresponds to a cluster where FAU is native, while the M- and H- attractors display MFI and HEU as native, respectively.

With this outcome, we calculate the mean radius $<R_k>$ to the instances centroid of the k^{th} attractor and determine the corresponding 9-D number density $\rho_k = N_k/(CR^9)$, where N_k is number of instances of the k^{th} attractor, $C = 2^5\pi^4/945$ and $k = 1$ through 4. The M-attractor is the densest cluster and L-attractor is the least dense cluster. Densities of the other two attractors are similar and lay in-between the attractors with extreme densities. Having distances between instances is useful for a visual representation of the attractor points in 9-D space through calculation of radial distribution functions. Radial distributions provide a means of determining the probability of finding instances in spherical shells around each given instance. Results of the semisupervised case are shown in Figure 2 depicting the ten radial distribution functions generated within the four attractors and between them (normalized by the sum of distances). The distribution functions give a hint of the cluster shape in the 9-D space. For example, g_L indicates that this attractor is elongated and has a neck, while g_F depicts a more concentrated spherical shape and g_M clearly explains the fact that this cluster is the most dense of the four. The inter-attractor distribution functions show the absence of small distances, which is to be expected if the clusters are well separated in the 9-D space. Figure 3 > gives a 3-D histogram on the scaled plane of two selected features (V_1, V_c)

Table 1. The four framework attractors obtained through clustering and label specificity. Mean and s.d. are obtained over ten datasets.

	Label	Unsupervised specificity	Semisupervised specificity
L-attractor	LTA	1.00 ± 0.0	1.00 ± 0.0
	CHA	0.94 ± 0.10	0.96 ± 0.06
F-attractor	FAU	1.00 ± 0.0	1.00 ± 0.0
	KFI	1.00 ± 0.0	1.00 ± 0.0
	RHO	0.84 ± 0.11	1.00 ± 0.0
	CHA		0.04 ± 0.06
	EDI	0.05 ± 0.0	
M-attractor	MFI	1.00 ± 0.0	1.00 ± 0.0
	MOR	0.98 ± 0.05	0.99 ± 0.04
	FER	0.61 ± 0.58	0.97 ± 0.05
	HEU		0.01 ± 0.04
H-attractor	HEU	1.00 ± 0.0	0.99 ± 0.04
	LAU	1.00 ± 0.0	1.00 ± 0.0
	LTL	1.00 ± 0.0	1.00 ± 0.0
	NAT	1.00 ± 0.0	1.00 ± 0.0
	EDI	0.95 ± 0.0	1.00 ± 0.0
	FER	0.39 ± 0.58	0.03 ± 0.05
	RHO	0.16 ± 0.11	
	CHA	0.06 ± 0.10	
	MOR	0.02 ± 0.05	0.01 ± 0.04

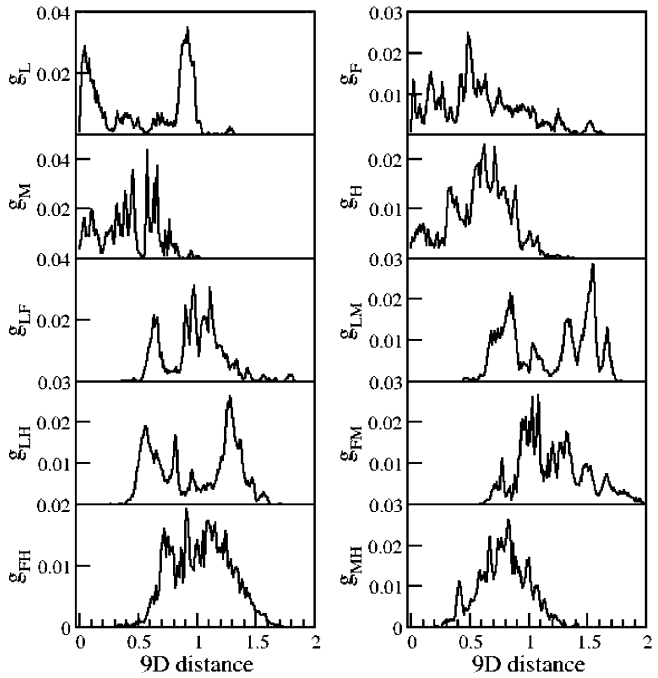

Figure 2. Normalized pair distribution functions of the 9-D distances.

with the height being the frequency of occurrence in ten data sets. Although the histogram gives a partial pictorial representation of the four attractors, it points out to the existence of the grouping of instances.

A verification on the goodness of the hybrid hierarchical clustering is to calculate the total point scatter P_S and the loss function W (in-cluster scatter) defined as

$$P_S = \sum_{i=1}^{246} \sum_{j>i}^{247} d(x_i, x_j) \quad (3)$$

which is constant for each of the ten data sets. The loss function is

$$W = \sum_{k=1}^{4} \sum_{i=1}^{N_k-1} \sum_{j>i}^{N_k} d(x_i, x_j) \quad (4)$$

where $d(x_i, x_j)$ are the squared distances defined in Equation 2 and N_k is the number of instances in each attractor. In a good clustering mechanism, W needs to be small when compared to the between-cluster function $B = P_S - W$. The ratio W/B is 0.15 and averages and standard deviation (s.d.) over the ten sets are $W = 3633 \pm 701$, $B = 23868 \pm 1146$ for the unsupervised case. The semisupervised case yields very similar results: $W/B = 0.14$, $W = 3297 \pm 291$, $B = 24205 \pm 964$. As expected, W is about one order of magnitude smaller than B. Other dissimilarity metrics have been tested in clustering studies.[17] For example, two different metrics are based on Manhattan distances (also called taxicab distances) and Chebychev distances (also referred to as chessboard distances).[18] The two metrics give dual polyhedra when a sphere is built from them. When Manhattan distances are used in Equations 3–4, the ratio W/B is 0.13 for the unsupervised case and 0.11 for the semisupervised case. Use of the Chebychev distances yields $W/B = 0.18$ and 0.17 for the unsupervised and semisupervised cases, respectively. These results are quite close, reinforcing the validity of the definition of the four attractors.

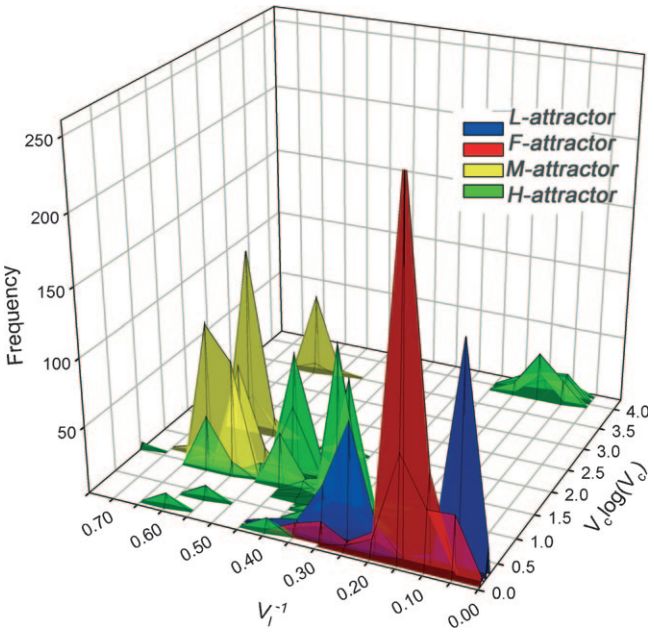

Figure 3. Visual representation of the four attractors in a scaled plane of two features.

Different measures of the precision with which instances are grouped into clusters exist in machine learning. A popular measure is the cluster impurity index[19] that becomes zero for pure clusters. Impurity in the unsupervised case is 4.7%, which improves significantly to 0.70% in the semisupervised case. Another measure is the modified-impurity-Gini,[1] which is 8.57% for the unsupervised case and decreases to 1.37% for the semisupervised case.

Our data analysis has proven that four attractors are clearly defined. An alternative way to verify the clustering result is to build a supervised classification model with the attractors as classes. The classification method Random Forest[20] with 100 trees and ten-fold cross validation was selected. Ten classification models were created, one model with each of the ten data sets containing 247 instances each and the 9-D feature vector. The purpose of each model is to classify unknown instances into one of four classes. Once the model is saved, then all other instances in the 1034 set not used in building the model are classified. Results yield an accuracy (percentage of correctly classified instances) of 98.95 ± 0.68% for the unsupervised case and 99.49 ± 0.4% for the semisupervised case showing an

almost perfect classification. Another measure of the classification is the out-of-bag error OOB,[2] which is 0.15 ± 0.006 for both cases.

To ensure that the classification results are not fortuitous, ten random models were built in which the assignment of instances belonging to any one of the four attractors was given at random. Results are $30.20 \pm 7.40\%$ accuracy for the unsupervised case and $28.58 \pm 6.37\%$ accuracy for the semisupervised case, which coincide with the expected random accuracy of 100/4%.

4 Conclusions

Our informatics approach of unsupervised statistical learning has identified four super-classes of zeolite framework types showing that the zeolite structures form clusters around four core framework types: LTA, FAU, MFI and the combination of EDI, HEU, LTL and LAU. Clustering occurs in a nine-dimensional space of attributes that reflect topological, chemical and physical properties of zeolite crystalline structures. The implemented machine learning approach relies on top-down clustering and the EM method. Our clustering model is trained and tested on ten partially independent data sets from the ICSD. The four clusters are named L-, F-, M- and H- attractors and constitute four global classes into which newly synthesized or mathematically predicted zeolite structures could be identified. The proposed new structural classification is useful for speeding up the process in material design of porous materials.

Acknowledgements

This work was supported by the *National Science Foundation Grant CHE-0626111*. The TeraGrid resources provided by the *Pittsburgh Supercomputing Center* are acknowledged. Authors are grateful to the Standard Reference Data Program of *NIST* for making available the zeolite dataset from the FIZ/NIST ICSD in ascii format.

References

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, **2009**.
[2] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco, CA, **2004**.
[3] A. P. Dempster, N. M. Laird, D. B. Rubin, *J. Roy. Stat. Soc. B* **1977**, *39*, 1–38.
[4] K. Pearson, *Philosoph. Magazine*, **1901**, 2, 559–572.
[5] A. Blum, T. Mitchell, in *Proc. 11th Ann. Conf. Computational Learning Theory*, Madison, WI, **1998**, pp. 92–100.
[6] Committee on the Fundamentals of Computer Science: Challenges and Opportunities, National Research Council, *Computer Science: Reflections on the Field, Reflections from the Field*, The National Academies Press, Washington DC, **2004**.
[7] H. Steinhaus, *Bull. Acad. Polon. Sci.* **1956**, *4*, 801–804.
[8] J. MacQueen, in *Proc. 5th Berkeley Symp. Math. Stat. Probability, Vol. I, Statistics* (Eds: L. M. Le Cam, J. Neyman), University of California Press, Berkeley, CA, **1967**, pp. 281–297.
[9] S. P. Lloyd , *IEEE Trans. Information Theory* **1982**, *28*, 129–137.
[10] S. Yang, M. Lach-hab, I. I. Vaisman, E. Blaisten-Barojas, in *Proc. 2009 Int. Conf. Artificial Intelligence* (Eds: H. R. Arabnia, D. de la Fuente, J. A. Olivas), CSREA, Las Vegas, NV, **2009**, pp. 340–344.
[11] *FIZ/NIST Inorganic Crystal Structure Database*, http://www.nist.gov/srd/nist84.htm, **2007**.
[12] *IZA-SC database of ideal zeolite structures*, http://www.iza-structure.org/databases, **2009**.
[13] S. Yang, M. Lach-hab, I. I. Vaisman, E. Blaisten-Barojas, *LNCS* **2009**, *5545*, 160–168.
[14] S. Yang, M. Lach-hab, I. I. Vaisman, E. Blaisten-Barojas, *J. Phys. Chem. C* **2009**, 113, 21721–21725.
[15] D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, E. Blaisten-Barojas, *Microporous Mesoporous Mater.* **2009**, *117*, 339–349.
[16] *Weka 3: Data Mining Software in Java*, http://www.cs.waikato.ac.nz/ml/weka, **2009**.
[17] A. Al Kalifa, M. Haranczyk, J. Holliday, *J. Chem. Inf. Model.* **2009**, *49*, 1193–1201.
[18] *Handbook of Massive Data Sets* (Eds: J. M. Abello, P. M. Pardalos, M. G. C. Resende), Springer, New York, **2002**.
[19] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco, **2005**.
[20] L. Breiman, *Machine Learning*, **2001**, *45*, 5–32.

Received: November 23, 2009
Accepted: February 27, 2010
Published online: March 29, 2010