Data mining for materials: Computational experiments with AB compounds

Yousef Saad,[1] Da Gao,[1,*] Thanh Ngo,[1] Scotty Bobbitt,[2] James R. Chelikowsky,[2] and Wanda Andreoni[3,4]

[1]*Department of Computer Science & Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA*

[2]*Department of Chemical Engineering, and Center for Computational Materials, Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, Texas 78712, USA*

[3]*Centre Européen de Calcul Atomique et Moleculaire (CECAM), École Polytechnique Fédérale de Lausanne, Switzerland*

[4]*Institut de Théories des Phénomènes Physiques, École Polytechnique Fédérale de Lausanne, Switzerland*

(Received 28 November 2011; published 6 March 2012)

Machine learning is a broad discipline that comprises a variety of techniques for extracting meaningful information and patterns from data. It draws on knowledge and "know-how" from various scientific areas such as statistics, graph theory, linear algebra, databases, mathematics, and computer science. Recently, materials scientists have begun to explore data mining ideas for discovery in materials. In this paper we explore the power of these methods for studying binary compounds that are well characterized and are often used as a test bed. By mining properties of the constituent atoms, three materials research relevant tasks, namely, separation of a number of compounds into subsets in terms of their crystal structure, grouping of an unknown compound into the most characteristically similar peers (in one instance, 100% accuracy is achieved), and specific property prediction (the melting point), are explored.

DOI: 10.1103/PhysRevB.85.104104 PACS number(s): 61.50.Ah, 61.66.Fn, 64.70.dj, 89.20.Ff

I. INTRODUCTION

Data mining is a broad discipline that develops methods and tools to extract meaningful information and patterns from data. It draws on knowledge and "know-how" from various scientific areas such as statistics, graph theory, linear algebra, databases, mathematics, and computer science. With the emergence of the information era, the importance of these techniques has increased dramatically in information-related applications including commerce, finance, and criminology, to cite just a few. Data mining has also become an essential tool in the area of genomics, whose primary technique involves routinely sifting through millions of genes to discover similarities or patterns among them.

Materials scientists have begun to explore data mining ideas for the selection of materials in applications that range from photovoltaics to thermoelectrics to catalysts.[1,2] The following section gives a brief overview of a few basic techniques used in data mining, in part to define terminology. Whenever possible, an attempt is made to give examples from materials where the techniques can be applicable.

II. BASIC DATA MINING TECHNIQUES

Among the many problems that are tackled by data mining, two are of primary importance. One is "unsupervised clustering," which is the task of finding subsets of the data such that items from the same subset are most similar and items from distinct subsets are most dissimilar. The second is classification (predictive modeling, supervised learning), whereby we are given a few distinct sets that are labeled (e.g., samples of handwritten digits labeled from 0 to 9), and when a new sample is presented to us we must determine to which of the sets is it most likely to belong. For the example of handwritten digits this is the problem of recognizing a digit given a data set of many labeled samples of already deciphered digits available (called a training set).

In order to perform these tasks, it is common to first process the given data set (e.g., a database of handwritten digits as represented by the values of the pixels) in order to reduce its dimension, i.e., to find a data set of much lower dimension than the original one but that preserves its main features. What is sometimes misunderstood is that *this dimension reduction step is not done for the sole purpose of reducing cost but mainly to reduce the effect of noise and in order to extract the main features from the data.*

A. Dimension reduction and principal-components analysis (PCA)

Two distinct types of methods have been proposed for dimension reduction. The first class of methods which can be termed "projective" includes all linear methods whereby the data matrix is explicitly transformed into a low-dimensional approximation. These projective methods find an explicit linear transformation to perform the reduction, i.e., they find an $m \times d$ matrix V and express the reduced dimension data as $Y = V^T X$. This class of methods comprises the standard PCA, the locality preserving projection,[3] orthogonal neighborhood preserving projections (ONPPs),[4,5] and variants of these.

The second class of methods that do not rely on explicit projections and are inherently nonlinear[6] find directly the low-dimensional data matrix Y, by simply imposing that a certain locality or affinity between nearby points be preserved. Many of these methods utilize weighted graphs to represent the local geometry in a high-dimensional space, which they try to preserve. As an example, the locally linear embedding technique starts by defining a graph that expresses every point of the original data as an approximate convex combination of its immediate neighbors. Then it asks the question: How can we map these data points into a low-dimensional space (d coordinates instead of m, with $d \ll m$) in such a way that this graph is preserved as best as possible. This is illustrated on the right in Fig. 1.

These types of dimension reduction methods can be extended to supervised analogs, i.e., to situations where each data point is associated with a class label. The class labels are

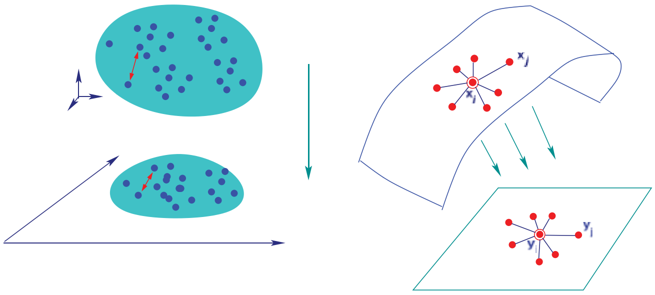

FIG. 1. (Color online) Dimensionality reduction techniques. Left: General method. Right: Graph-preserving method.

then taken into account when performing the reduction step. In the case of graph-based methods, this can be simply achieved by defining the neighbors of an arbitrary vertex i in the graph to be all the vertices which share the label i. Techniques based on this approach can be very powerful for face recognition; see, e.g., Refs. 5, 7, and 8.

Consider the PCA approach for dimension reduction. The primary assumption that makes PCA useful in this context is that there is some underlying low dimension of the high-dimensional data, which represents the most significant features of the data. If we are able to discover this space, we can perform whatever analysis we wish with fewer parameters. In PCA, this space is obtained via the singular value decomposition.[9,10] Specifically, let us denote by $\bar{X}$ the matrix of zero recentered data, i.e., each column is $\bar{x}_i = x_i - \mu$, where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of X. In PCA, an orthogonal matrix V is computed which will map the data so as to maximize the variance of the projected data in the d-dimensional space. As it turns out, the column vectors in this matrix V are the left singular vectors of $\bar{X}$, associated with the largest d singular values:

$$[\bar{X}\bar{X}]^T v_i = \lambda_i v_i, \quad i = 1,2,\ldots,d. \tag{1}$$

The matrix Y, corresponding to the projected (low-dimensional) data, is then given by $Y = V^T\bar{X}$.

Though materials informatics is a relatively new specialty, the use of databases in materials dates to the 1960s, with the emergence of extensive data sets. The key to "soft" design of materials, i.e., design without physical experimentation, is to keep the number of computational tests with materials to a minimum. This means that a search must be performed to select good candidate materials, which can be studied in more detail by solving the electronic structure problem for the properties of interest.

A recent example of this type comes from Curtarolo *et al.*,[11] who demonstrate an interesting application of PCA for the task of predicting structural energies of crystals with the help of the CRYSMET database. For 55 alloys, they form an array with 55 columns (for each alloy) and 114 rows (1 for each possible crystal structure). The structural energies, determined by density functional theory calculations, are correlated and these correlations are unraveled by PCA. With an rms error of 50 meV, only 9 dimensions are required of 114. The implication is that *it is not necessary to perform 114 experiments for a new alloy, but only 9*; the others can then be deduced from the correlation.

B. Unsupervised learning

In unsupervised learning, one is given a data set (refer to the example in Sec. III) and is then asked to find characteristics of the set using only the data at hand. For example, we may be interested in partitioning the set into distinct subsets. A number of techniques are used for this purpose and we refer the reader to standard textbooks, e.g., Refs. 12–14.

C. Supervised learning

Supervised learning tools are at the basis of pattern recognition. A prototypical application is that of "face recognition" or "digit recognition" (mentioned above). In face recognition, we are given a database of photographs picturing c known individuals (say, 20 photos for each of 100 known, i.e., labeled, persons). We are then presented with a test photo of an unknown person and would like to know if this person is 1 of the 100 labeled individuals. A simple comparison based on the array of pixels will generally perform very poorly. PCA is satisfactory in some cases, but graph-based methods such as ONPPs[5] perform quite well for applications where images are involved.

In this context, a number of powerful techniques have been developed in the literature to "classify" data, i.e., to find its class. Linear classifiers such as linear discriminant analysis and Fisher methods provide ways to optimally separate data into classes.

In the context of materials, one may apply this to guess the "class" of a given material. For example, we can consider a database of known (i.e., previously studied) compounds, which can be labeled "photovoltaic," and we now consider a given ternary material not studied before. From knowledge of its constituent atoms, and from known structures, we would like to know if it is likely to be a member of the photovoltaic class. When a good candidate material is identified, a full-fledged electronic structure calculation, e.g., one based on density functional theory, can be performed and the resulting data will then be added to the database. The method by which the material has been correctly or incorrectly classified will be updated according to the result. This feedback loop to improve the classification model is called "learning."

A major part of supervised learning is concerned with building "classifiers," which will help determine whether or not a given new material has a certain property. For example, is the material in the multiferroic class or not? If the illustration in Fig. 2 represents a cloud of materials in some high-dimensional space, the simplest form of classifier is just a hyperplane, which will tend to best separate the multiferroic materials from the others. The picture may deceive one into believing that this is an easy task in this particular situation. However, two classes may not be easy to separate in a general case, especially in high dimensions. High dimensionality is one reason why "kernels" are commonly used in this context.[13] The use of kernels amounts to simply changing inner products so as to alter the notion of lengths.

D. Property prediction

A rather common question in materials is whether or not it is possible to predict a value associated with some physical

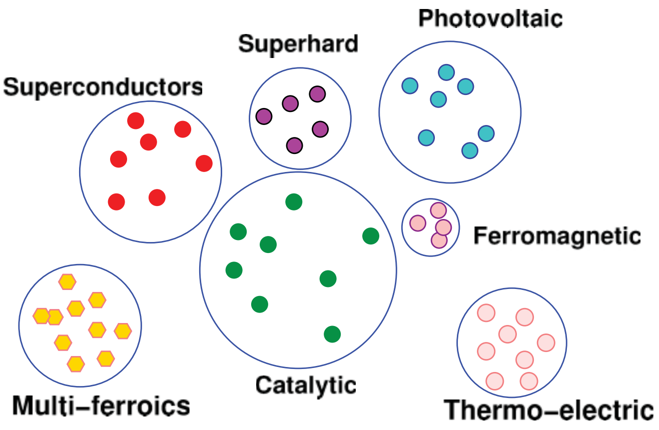

FIG. 2. (Color online) Classification of materials.

properties of a compound, e.g., its melting point. Ideally, these should be predictable from properties of the constituent atoms. The capability to predict a property value with a certain degree of accuracy is another important application of data mining techniques in materials research. Unlike supervised learning or unsupervised learning, in most cases, data mining techniques are combined with statistical regression methods to generate a numerical physical property value of an unknown material. The ultimate goal is to discover the genuine function that can precisely describe the correlation between the variable to be predicted and other already known parameters. The regression part works by finding the best fit to a set of points. Take linear regression as an example. The best fit is achieved by finding the minimum value of the squared residuals, leading to what is known as the least-squares method. At the same time, data mining techniques can efficiently extract the main features from the data and reduce the effect of noise. As a result, the unique combination of regression and data mining ideas may provide a powerful mechanism for predicting numerical values of materials properties.

III. UNSUPERVISED LEARNING EXPERIMENT

We illustrate “unsupervised learning” by considering a well-known family of crystal structures. These are binary octet crystals whose composition is $A^N B^{8-N}$, where N refers to the number of valence electrons. This family of crystals includes technologically important semiconductors such as Si, Ge, GaAs, GaN, and ZnO. There are approximately 80 members of this crystal family, which condense primarily in graphite, diamond (D), zincblende (Z), wurtzite (W), rocksalt (R), and cesium chloride structures.

The separation of these structures into distinct classes is difficult and has existed as a problem in the literature for over 50 years.[16–21] Ordinary chemical coordinates such as size and electronegativity will not result in topologically distinct regimes.[22]

Figure 3 illustrates one of the most successful structural maps for this family. The separation between structural types is nearly exact. Of special note is the separation between the Z and the W structures. These two family types often differ by only ~0.01 eV/atom, as the Z and W structures are nearly identical in terms of local order. They differ only in the third nearest neighbor. The chemical coordinates (r_σ, r_π) employed in Fig. 3 were based on orbital radii determined from model pseudopotentials fit to spectroscopic data.[15,23,24] In particular, the orbital radii are based on pseudopotential description of the free ion. For example, the silicon radii are constructed from considering Si^{+3} ions, i.e., one electron moving in the field of the silicon ion core. The model pseudopotential is taken to be

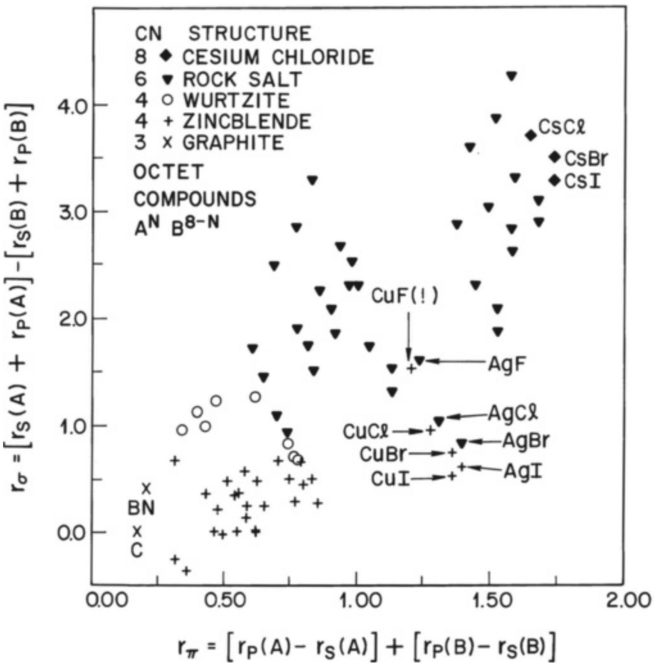

FIG. 3. Structural map for binary octet crystals. The coordination number (CN) is indicated for each structural grouping. The chemical coordinates (r_σ, r_π) are combinations of orbital radii as defined in Ref. 15. This mapping of these compounds in two dimensions with the particular coordinates used in Ref. 15 reveals a good clustering of the structures.

$$V(r) = -\frac{Z_v}{r} + \sum_{l=0}^{Z_v} \frac{\hat{l}(\hat{l}+1) - l(l+1)}{2r^2}\mathcal{P}_l. \tag{2}$$

Here, Z_v is the number of valence electrons, $\mathcal{P}_l$ is a projection operator, which projects the lth component of the angular momentum, and $\hat{l}$ is an l -dependent parameter. Atomic units $(\hbar = e = m)$ are used. This potential replicates only the valence states. A key advantage of this potential is that it has an analytic solution for the energy levels of the ion. The energy levels can be written as

$$E_{n,l} = \frac{-Z_v^2}{2(n+\hat{l}-l)^2}. \tag{3}$$

The energy levels can be interpreted as Rydberg levels, with an l-dependent defect given by $\hat{l}$. Orbital radii can be defined by finding the classical turning points, $V(r_l) = 0$, or the radial maximum of the wave functions arising from this potential[15,23] as 2 differ by a factor of 2. The turning points are probably more physical, but traditionally the radii are defined by the maximum of the wave function and are given by

$$r_l = \hat{l}(\hat{l}+1)/Z_v. \tag{4}$$

While this pseudopotential is not particularly good for calculations, e.g., it possesses a divergent potential in the core region and the wave functions are not similar to those expected for an all-electron potential, this potential is good for extracting the orbital deviations from a hydrogenic atom and thus characterizing the chemical nature of the ion core. The orbital radii are determined once $E_{n,l}$ is known. The energy levels can be determined experimentally from spectroscopic data, but the use of spectroscopic data has some obvious disadvantages. Consider an atom like fluorine. To define the orbital radii for fluorine, we would need to consider an F^{+6} ion, which is extraordinarily difficult to create and measure. In the original work,[15,23,24] the radii for such cases were estimated by extrapolation from known values of the energy levels.

Here we have decided to update the radii by considering theoretical calculations for the energy levels and avoid the use of spectroscopic data. We use density functional theory to determine the total energy to remove an electron from the ion of interest. For example, we consider a Si^{+3} ion with a configuration of $3s^1 3p^0$ for the s state and $3s^0 3p^1$ for the p state. We determine the total energy of the ion with these configurations and then subtract the energy of the ion core. We employ the local density approximation, which is known to be very accurate for ionization energies of neutral and positively charged atoms.[25,26] For heavy atoms such as cadmium and cesium, we include relativistic effects, which tend to result in small values of r_s. The new set of radii produces a plot very similar to the one illustrate in Fig. 3.

The two-dimensional (2D) mapping used in this example of octet compounds is identical to what is usually done in dimension reduction for visualizing complex data. Figure 3 shows the compound CuF, which was thought to exist in the form of a Z structure as noted in another publication.[15] The mapping revealed that this hypothetical compound is surrounded by crystals in the R structure. Further research showed that the CuF compound does not actually exist as suggested by the 2D mapping.[15]

The 2D mapping in this example was performed by a judicious change of coordinates, exploiting physical intuition. One question that may be asked is whether or not a similar mapping can be discovered in some systematic way. If we restrict the mapping to be linear, then the answer depends on what "features" are included in the data.

In our experiment, we use only the following information from each of the two constituent atoms:

(1) the number of valence electrons;

(2) the ionization energies of the s and p states of the ion core; and

(3) the radii for the s and p states as determined from model potentials, which are also listed in Table I.

The total number of valence electrons is eight for the compounds considered, so there is some redundancy in these data. Since we are considering two atoms, we will normally have 10 features available for each compound, or 9 actually, because the number of valence electrons for the B atom can be obtained from the first. With nine features the data are still somewhat redundant, in part because some elements repeatedly appear in different compounds.

The data set we consider is basically the same as before and consists of 67 compounds. In this study, we did drop the copper and silver halides, as we wanted to restrict our study to compounds made with simple metals and avoid complications associated with d valence states. For example, should we consider the d states in copper as part of the valence shell or not? We classify these compounds into six structures—Z, W, D, R, and "dual structures"—where the ground-state structures are borderline between two phases: Z and W (ZW) and W and R (WR). Z and W are particularly difficult cases because the structures differ only past second nearest neighbors. Such degenerate structures can occur at ambient temperature and pressure. As an example, ZnS can occur as a Z structure or as a W structure. To accommodate such situations we would label ZnS as belonging to the ZW class.

In addition, the raw use of the number of valence electrons leads to some difficulties, as these numbers are of a different scale from the others. As a result, for each atom we simply use

TABLE I. List of radii used in the present work. Radii were based on Eq. (4) using Kohn-Sham energy levels and are given in atomic units.

Element	r_s	r_p
Li	0.99	1.93
Be	0.66	0.96
B	0.49	0.64
C	0.40	0.48
N	0.33	0.39
O	0.28	0.33
F	0.25	0.28
Na	1.01	2.35
Mg	0.86	1.42
Al	0.75	1.09
Si	0.66	0.88
P	0.59	0.75
S	0.54	0.66
Cl	0.49	0.59
K	1.34	2.68
Ca	1.22	1.84
Cu	0.37	1.48
Zn	0.62	1.17
Ga	0.65	1.01
Ge	0.64	0.90
As	0.62	0.82
Se	0.59	0.75
Br	0.57	0.70
Rb	1.44	2.86
Sr	1.36	2.05
Ag	0.47	1.58
Cd	0.67	1.26
In	0.78	1.15
Sn	0.78	1.06
Sb	0.76	0.98
Te	0.74	0.92
I	0.71	0.87
Cs	1.66	3.08
Au	0.22	1.32
Ba	1.52	2.29
Tl	0.67	1.13
Hg	0.57	1.21
Pb	0.71	1.06
Bi	0.71	1.00

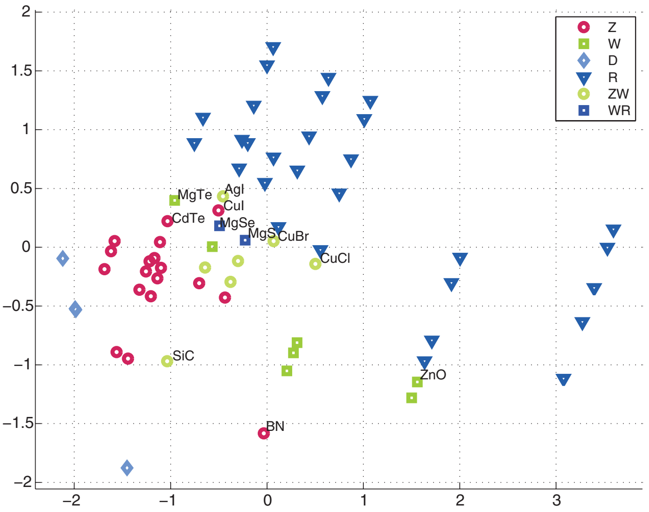

FIG. 4. (Color online) PCA projection for 67 octet compounds.

the number of valence electrons to scale the data. Specifically, if Z_v is the number of valence electrons for a given atom, we use the following information:

(1) the energies of the s electron and the p electron scaled by $\sqrt{Z_v}$ and

(2) the radii of s-electron and p-electron orbitals.

With this we can produce the data matrix used for the clustering experiment. The matrix is of size 8×67. Each column corresponds to 1 of the 67 binary octets considered. The entries (rows) are simply the four features mentioned above for atom A followed by the same features for atom B. For elemental crystals, we simply repeat the information, essentially making the AB compound with $B == A$. We then use PCA (and other techniques) to project the data in two dimensions. This gives a 2×67 array, i.e., two coordinates for each octet. These two coordinates are used to plot the data in a 2D plane. The result is shown in Fig. 4. The dual-structure compounds ZW and WR are represented with the color of one structure and the shape of the other to facilitate interpretation. As can be seen, the R compounds are nicely separated from the other structures, as are the D structures.[27] For the sake of lightening the figure, only the labels of a few borderline crystals are shown.

IV. SUPERVISED LEARNING EXPERIMENT

This section illustrates what is commonly referred to as "supervised learning" in data mining. The problem at hand is to try to identify the unknown "class" of a given compound. This class can be a property such as "photovoltaic" or "superconductor," etc. It is a label we assign to an item. In the experiment described, the class is the structure of the compound, one of the six labels Z, W, D, R, ZW, and WR.

The problem setting is as follows. We have n compounds $c_1, c_2, \ldots, c_n$ whose classes $s_1, s_2, \ldots s_n$ are known. This set is commonly referred to as the "training set." We also have another compound called t (for "test"), whose label (structure) is unknown. The problem is to determine the class of t. To do so we need to use the information we have about these compounds. In the illustration below we are allowed to use the exact same information as in the previous section (nine entries in all for each compound). We describe three methods which are known for their simplicity.

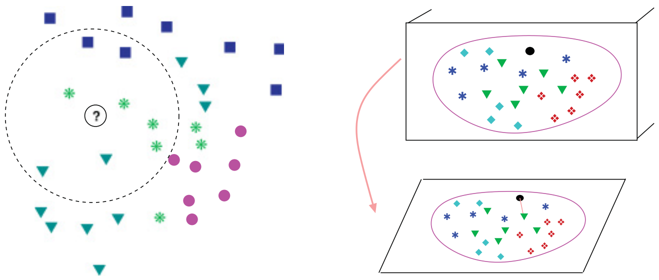

FIG. 5. (Color online) Left: KNN classification. Right: Classification by PCA projection.

The first method (KNN) uses a majority rule among k nearest neighbors. In this approach, illustrated in Fig. 5 (left), some distance between the test sample and all other compounds is evaluated and the classes of the k nearest neighbors (eight in the figure) are considered. We attribute to t the label of the predominant class among these k items. For the example in the figure the test sample will get the class "asterisk." The issues with this method are what is a good choice for k and what distance to use. In the experiment we only use $k = 5$ and the Euclidean norm distance.

The second approach is based on the observation made in the earlier section that PCA does an excellent job at reducing dimensionality. PCA can then be used for classification. We project everything in a low-dimensional space and determine the closest item to t in this low-dimensional space. The class assigned to t will be the class of this item. This is a common technique used in the area of pattern recognition, as, for example, when we try to recognize an individual in a photo (face recognition) by comparing a "test image" with pictures of a number of known individuals.

We describe a third method, which we refer to as ONPPs.[4] This method seeks an orthogonal mapping of a given data set so as to best preserve a certain affinity graph. The graph we use here is the one associated with the classes: any two compounds in the same class will be linked by an edge. This means that a class forms a "clique." We then associate a weight matrix W with this graph in which an entry w_{ij} has the value 0 if i and j are not in the same class and $1/|C|$ if they both belong to class C. (Note that $|C|$ is the cardinality of this class C.) The projection matrix V in ONPP is determined so that V is orthogonal ($V^T V = I$) and so that the projected data $Y = V^T X$ minimize the sum of $w_{ij} \| y_i - y_j \|$ over all pairs i, j. This encourages y_i and y_j to be close. After some algebraic manipulations, the optimization problem becomes

$$\min_{\substack{V \in \mathbb{R}^{m \times d} \\ V^T V = I}} \operatorname{Tr}\,[V^T X (I - W^T)(I - W) X^T V]. \tag{5}$$

Its solution is the basis of the eigenvectors associated with the d smallest eigenvalues of the eigenvalue problem:

$$X(I - W^T)(I - W)X^T u_i = \lambda u_i. \tag{6}$$

Then the projector V is $[u_1, u_2, \ldots, u_d]$ and results in the projected data $Y = V^T X$.

TABLE II. Recognition rate for three methods using the data in different ways.

	KNN	ONPP	PCA
Case 1	0.909	0.945	0.945
Case 2	0.945	0.945	1.000
Case 3	0.964	0.945	0.982
Case 4	0.909	0.964	0.964
Case 5	0.945	0.964	0.945
Case 6	0.964	0.964	0.945

TABLE III. Recognition details for case 6.

Compound	Structure	KNN	ONPP	PCA
BeO	W	W	W	W
LiF	R	R	R	R
BP	Z	Z	Z	Z
SiC	ZW	Z	Z	Z
BeS	Z	ZW	ZW	Z
AlN	W	W	W	W
LiCl	R	R	R	R
MgO	R	R	R	W
NaF	R	R	R	R
BAs	Z	Z	Z	Z
AlP	Z	Z	Z	Z
MgS	WR	WR	WR	WR
BeSe	Z	ZW	ZW	Z
GaN	W	W	W	W
ZnO	W	W	W	W
LiBr	R	R	R	R
NaCl	R	R	R	R
CaO	R	R	R	R
KF	R	R	R	R
BeTe	Z	Z	Z	Z
AlAs	Z	Z	Z	Z
GaP	Z	Z	Z	Z
ZnS	ZW	Z	Z	Z
MgSe	WR	WR	WR	WR
LiI	R	R	R	R
CdO	R	W	W	W
InN	W	W	W	W
CaS	R	R	R	R
NaBr	R	R	R	R
KCl	R	R	R	R
SrO	R	R	R	R
RbF	R	R	R	R
AlSb	Z	Z	Z	Z
GaAs	Z	Z	Z	Z
InP	Z	Z	Z	Z
MgTe	W	Z	Z	WR
ZnSe	ZW	Z	Z	Z
CdS	ZW	ZW	ZW	Z
NaI	R	R	R	R
CaSe	R	R	R	R
SrS	R	R	R	R
KBr	R	R	R	R
RbCl	R	R	R	R
GaSb	Z	Z	Z	Z
InAs	Z	Z	Z	Z
ZnTe	Z	Z	Z	Z
CdSe	W	ZW	ZW	Z
CaTe	R	R	R	R
KI	R	R	R	R
SrSe	R	R	R	R
RbBr	R	R	R	R
InSb	Z	Z	Z	Z
CdTe	Z	Z	Z	Z
SrTe	R	R	R	R
RbI	R	R	R	R

The data set we consider consists of the same set as before except that we have removed the elemental crystals for the moment because they are isostructural with the Z structure; i.e., if one ignores the difference in atomic species, the Z and D structures are identical. We have also removed all the Cu and Ag crystal structures as mentioned before, as well as BN, since it alone occurs in a graphite structure once C is removed.

This leaves us with a set of 55 compounds. We perform a "leave-one-out" experiment in which we take each of the 55 compounds in turn and pretend we do not know its structure. We then try to guess its structure by correlating it with the other 54 compounds. The average precision, i.e., recognition rate of the process, is then computed for all 55 cases. This is the mean number of times (out of 55) that the procedure guessed the correct structure and it is computed for each method separately. For the situations where a compound has a dual structure, we decided to rate as correct any outcome where at least one label of the two matches. For example, if the system returns a WR for an R, we rate the outcome as correct. Similarly, the outcome is rated correct in the case when a WR is returned for a W.

Table II presents the results for the following cases.

Case 1. For each atom, use features 1:5 for atom A and 2:5 for atom B. No scaling is applied.

Case 2. For each atom, use features 2:5 for atom A and atom B. Scale features 2 to 4 (s and p energies and s radius) by $\sqrt{z}$.

Case 3. For each atom, use features 1:5 for atom A and 2:5 for atom B. Scale features 2 and 3 (s and p energies) by $\sqrt{z}$.

Since ONPP and PCA are projection-type methods, we can use two distances when trying to determine a class. We can elect to compare VV^Tt with x_i by measuring $\|VV^Tt - x_i\|$ or we can work in the V space by comparing V^Tt with V^Tx_i, i.e., by measuring $\|V^Tt - V^Tx_i\|$. Cases 1–3 use the former measure. Cases 4–6 are identical to cases 1–3 but use the second measures, i.e., those based on $\|V^Tt - V^Tx_i\|$. These are different distances when the projector does not project x_i exactly, i.e., when $VV^Tx_i \neq x_i$. Table III details the structure recognition, in case 6, for all 55 compounds.

Looking at Table II we note that even KNN, the simplest method, achieves a recognition of at least 94.5% in four of the six tests. The other two methods easily achieve recognition rates of 96.4% and higher (2 errors of 55). In one instance of PCA (test 2), 100% accuracy is achieved, although this is a rather contrived situation shown here only to illustrate the possibility of getting 100% accuracy. One compound that is not easily recognized by any of the procedures is MgTe. This is a W, identified incorrectly by KNN and by ONPP as a Z in all six tests. It was labeled WR by PCA in cases 2, 3, and 6 and Z in all other tests. CdO (an R) also gave difficulties. It was incorrectly labeled as W by KNN in all six cases, by

ONPP in three of the six cases, and by PCA in two of the six tests.

V. PROPERTY PREDICTION EXPERIMENT

In this section we explore the melting point of 44 *AB* suboctet compounds—following an experiment performed in the paper[15] mentioned earlier. *AB* suboctet compounds are composed of simple metals and metalloids as cations and do not contain any transition metals; the number of valence electrons for the two components is less than eight, e.g., MgAu, NaIn, and LiAl. As discussed for the previous supervised learning experiment, we perform a "leave-one-out" experiment. Experimental melting points for this set of 44 compounds are available. By removing 1 of them, we are left with 43 and can use these data to perform a (linear) regression. The melting point is expressed as a linear combination of a number of selected features, such as the s radius and p radius of each of A and B, the number of valence electrons of atom A, the number of valence electrons of atom B, and so on.

A common method used for regression is simply the least-squares approach. However, in the presence of experimental data and ill conditioning, it is often the case that regularization must be used. Tikhonov regularization[28,29] has been applied for this test. In a standard regression analysis, we solve a least-squares problem, min $\|Xa - b\|_2$, where b are the measured values for each of the m samples, $\|.\|_2$ is the Euclidean norm, the columns in X represent variables evaluated for each of the m samples, and a is the sought coefficient vector, which determines how the variables are (optimally) combined to yield the result b. The solution to the problem is $a = X^\dagger b$, where $X^\dagger$ represents the pseudoinverse of X. In Tikhonov regulation an approximate optimal solution is found in the form $a = (X^T X + \tau I)^{-1} X^T b$, where τ is a regularization parameter. In our study we first normalize the data matrix X by scaling its rows by their 2-norms. The regularization parameter used is $\tau = 0.135$.

A combination of 16 features for each suboctet binary compound, namely, 8 features for each constituent atom, A and B, have been selected for the melting point prediction. These eight features for each atom are (1) the number of valence electrons, (2) the radius for the s states as determined from model potentials, (3) the radius for the p states as determined from model potentials, (4) the electron negativity, (5) the boiling point, (6) the first ionization potential, (7) the heat of vaporization, and (8) the atomic number. The radii for both the s states and the p states are listed in Table I. The electronegativity is the Pauling electronegativity.[30] The atomic number and the number of valence electrons are adopted from the periodic table published by the National Institute of Standards and Technology. Values of the other three features, namely, the boiling point, the first ionization potential, and the heat of vaporization, are listed in Table IV.

The results are reported in Table V and they are also shown in Fig. 6. The relative error in Table V is defined as the absolute difference between the experimental and the predicted melting point divided by the experimental melting point. The median relative error of our predictions is 0.128, which means that half of the predictions have a relative error that is less than 12.8%. The boundaries for predictions with a relative error of 15% are also plotted in Fig. 6. There are six compounds whose prediction relative error is above 25%, namely, 86% of the predictions have a relative error that is less than 25%. Only one compound's melting point is predicted with a relative error that is slightly above 30%. This compound is CaAg. Considering the fact that the experimental melting points range from 578 to 1573 K, the theoretical difficulty of a clear description and understanding of the melting point, and the relatively simple linear regression method used in the study, this performance is understandable. All it says is that it may be possible to use data mining to predict, within a moderate error (say, less than 15%), a value associated with some physical properties of a material from properties of the constituent atoms.

TABLE IV. List of the boiling points, first ionization potentials, and heats of vaporization used in the present work.

Element	Boiling point (K)	First ionization potential (eV)	Heat of vaporization (kJ/mol)
Ca	1757	6.11	154
Ag	2436	7.58	251
Ba	2171	5.21	142
Pb	2013	7.42	178
Ge	3103	7.90	331
Si	2628	8.15	384
Sn	2543	7.34	296
Sr	1657	5.70	144
Tl	1746	6.11	164
I	459	10.45	21
Cd	1038	8.99	100
Li	1615	5.39	146
Mg	1363	7.65	127
In	2346	5.79	232
Au	3080	9.23	334
Rb	961	4.18	72
Be	3243	9.32	292
Cu	2840	7.73	300
Hg	630	10.44	59
Al	2740	5.99	293
Ga	2676	6.00	259
Na	1156	5.14	97
Bi	1837	7.29	105
K	1032	4.34	80

It is remarkable that the anomalous behavior of MgAu disappears in the current study while it consistently appears in both the Bloch-Simons and the Mooser-Pearson analytical studies; this has been discussed extensively in Ref. 15. Such anomalous behavior for rather simple suboctet compounds hinders the practical application, as well as the further development, of both the Bloch-Simons and the Mooser-Pearson analytical models for melting point prediction. It may be attributed to the failure to incorporate correctly the *p*-*d* hybridization into the Bloch-Simons and Mooser-Pearson models. In some ways, the effects of the *p*-*d* hybridization are reduced via data mining. This may be partly due to the fact that more physical properties of the constituent atoms are incorporated in the prediction through data mining, and some of these properties may include *p*-*d* hybridization implicitly. Twenty-three properties of the constituent atoms, though less

TABLE V. Comparison of predicted and experimental melting points for suboctet compounds.

	Melting point (K)		
Compound	Experimental	Predicted	Relative error
CaAg	938	1266	0.349
BaPb	1123	1158	0.031
BaGe	1418	1450	0.023
CaGe	1573	1385	0.120
CaSi	1518	1378	0.092
CaSn	1260	1438	0.142
SrSi	1423	1551	0.090
SrGe	1438	1546	0.075
TlI	723	742	0.026
CdAg	1003	865	0.138
LiAg	1159	829	0.285
MgAg	1093	1090	0.003
ZnAg	963	1101	0.143
CdAu	900	1038	0.153
LiAu	918	1096	0.194
MgAu	1423	1114	0.217
RbAu	773	999	0.292
ZnAu	998	860	0.138
BeCu	1203	1413	0.174
CaCd	958	1089	0.137
CaTl	1243	930	0.252
CaHg	1234	979	0.206
SrCd	973	1046	0.075
ZnCu	1153	1040	0.098
LiHg	868	747	0.140
MgHg	900	982	0.091
LiPb	755	895	0.186
LiTl	783	743	0.051
MgTl	628	808	0.286
LiAl	991	1066	0.076
LiCd	822	811	0.013
LiGa	999	976	0.023
LiIn	910	896	0.016
NaIn	713	780	0.094
LiZn	753	917	0.217
NaTl	578	632	0.094
LiBi	878	895	0.020
NaBi	793	682	0.140
NaPb	641	714	0.114
KPb	843	830	0.016
KSn	1103	900	0.184
BaCd	854	1086	0.271
BaHg	1095	919	0.161
HgSn	1133	1016	0.103

than complete, have been studied and have led to the optimal feature set listed above. These features include the covalent radius, atomic mass, melting point, and ionic radius, to cite just a few. More importantly, the construction of the feature matrix via similar compounds of the target compound, as described below, provides a sound basis for the predictive inference.

Our melting point prediction algorithm works as follows: (1) Select one compound as a prediction target from the compound data set; (2) search the remaining compounds in the data set for similar compounds in terms of the angular momentum of the outermost orbital of the constituent atoms; (3) form the feature matrix using all compounds that are similar to the target compound; (4) use the feature matrix to perform a regression via Tikhonov regularization; (5) calculate the relative error and store the predicted melting point; (6) return to step 1 for the next unpredicted compound until all compounds have been predicted. Additional feature matrix formation mechanisms, such as atomic number and combination of compounds with only one constituent atom similar to the target compound, are also incorporated in our prediction process, forming a hierarchy of the feature matrix.

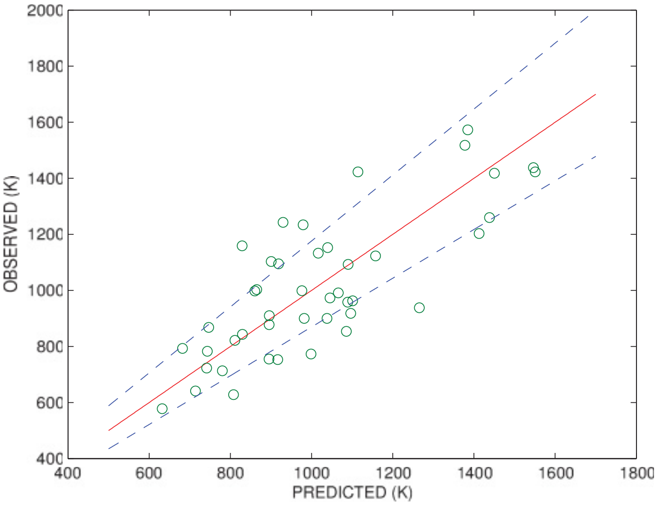

FIG. 6. (Color online) Experimental and predicted melting points for 44 suboctet binary compounds in degrees kelvin. The dashed (blue) line represents the boundaries for predictions with a relative error of 15%.

The worst predicted compound CaAg, with a relative error of 35%, is a compound with constituent atoms from s-block and d-block elements, respectively. Four of the six compounds, with a relative error greater than 25%, are compounds combined by atoms from both s-block and d-block elements too. In addition, 78% of the compounds, with a relative error greater than 20%, show the same characteristics, namely, one constituent atom is from the s-block elements while the other is from the d-block elements. Such a consistent pattern reveals that the lack of an accurate description for the d states may have a negative impact on predictions, regardless of the techniques applied. In Ref. 15, the d orbitals have been employed to explain the experimentally observed large melting point difference between MgAu and ZnAu. Our data mining experiments suggest that the complexity of the d orbitals is beyond the description of a single-parameter d-state radius. This is one reason why we omitted the coinage metals (Cu, Ag, and Au) when we considered the crystal structures.

In order to understand quantitatively the impact of each feature on the prediction accuracy, the sensitivity of features is also measured as follows. First, for the feature matrix $X \in \mathbb{R}^{m\times n}$, in which features are represented by columns while rows stand for compounds, the feature k for both atom A and atom B, namely, $X(:,k)$ and $X(:,k+8)$ is increased by the product of a uniformly distributed random number and the norm of the feature vector of the order of 10^{-8},

TABLE VI. Comparison of the sensitivities of different features.

Feature	Sensitivity
Number of valence electrons	809
Radius for s states	1650
Radius for p states	1057
Electron negativity	2384
Boiling point	2
First ionization potential	627
Heat of vaporization	17
Atomic number	92

represented as ϵ here. Consequently, the new feature values are $X(:,k) = X(:,k) + \epsilon$ and $X(:,k+8) = X(:,k+8) + \epsilon$ for both constituent elements of all compounds. Second, the new coefficient vector a_ϵ is then calculated according to $a_\epsilon = (X^T X + \tau I)^{-1} X^T b$, where τ is a regularization parameter. Finally, the vector norm of the difference between the new coefficient vector a_ϵ and the original coefficient vector a is divided by ϵ. Such a dimensionless ratio is calculated for all compounds, and its mean is assigned as the sensitivity of feature k, i.e., $\langle \|a_\epsilon - a\|/\epsilon \rangle$ where $\langle \ldots \rangle$ represents the sampling average. The above details the calculation of the sensitivity of feature k. Such a calculation is repeated for all features of the optimal feature set, as described previously, in order to obtain the sensitivity of all features. The results are listed in Table VI. Our results show that the electron negativity has the highest sensitivity value among the eight-feature set, which means that the change in the electron negativity will have the highest impact on the prediction accuracy. Furthermore, the similarity among compounds can be retrieved more via the electron negativity of the constituent elements than via any other single feature in the eight-feature set. In this similarity extraction mode, the eight features in the optimal set can be ranked, in descending order, as follows: (1) the electron negativity, (2) the radius for the s states, (3) the radius for the p states, (4) the number of valence electrons, (5) the first ionization potential, (6) the atomic numbers, (7) the heat of vaporization, and (8) the boiling point. It is interesting that the experimentally determined heat of vaporization and the boiling points are the lowest ranked. These features implicitly contain all possible attributes. Also, unless the structure of the melt is very different, the boiling points should contain essentially the same information as the heats of vaporization. As such, it is not surprising that the two features show similar behavior.

The melting point prediction study presented here suggests the possibility of promising applications of data mining techniques in materials property exploration. On the other hand, advancements in the physics, as well as insight into the nature, of materials, in particular, the electronic structure of materials, will greatly promote such data mining applications in materials research. In essence, the spirit of data mining applications in any field is the search for similarities that are relevant to the application goal. Unfortunately, the measurement of similarities among materials for a targeted material property is still at a nascent stage.

VI. CONCLUSIONS

The primary aim of this paper was to show how a few simple data mining techniques can be applied to answer a few specific questions on materials. In the first experiment, an "unsupervised learning" technique enabled us to separate 67 octet compounds into distinct classes according to their crystal structure through a PCA projection of the two constituent atoms' properties. In the second experiment, using a "supervised learning" technique, we were able to find the correct crystal structure of 55 compounds with an average success rate of 95%. In one instance of PCA, 100% accuracy was achieved, albeit with an *ad hoc* scheme. Finally, a simple form of regularized regression enabled us to predict the melting point of 44 suboctet compounds with a median relative error of 12.8%. This was achieved by mining a combination of 16 properties of the constituent atoms of each binary compound.

These preliminary results indicate that there is great potential for applying data mining techniques in materials science. This said, it is clear that more complex issues of materials science will lead to big challenges for data mining. On the bright side, there is much more to data mining than the basic techniques explored here. Once researchers gain a better understanding of the intrinsic nature of materials-related data, we will likely be in a much better position to deploy these methods for large data sets and extract much more meaningful information than was demonstrated in this article.

ACKNOWLEDGMENTS

The first three authors would like to acknowledge support from the National Science Foundation through Grant Nos. DMS-0810938 and DMR-0940218 and, also, would like to thank the Minnesota Supercomputing Institute for providing resources and an excellent environment for this research. S.B. and J.R.C. wish to acknowledge support from the National Science Foundation under Grant Nos. DMR-0941645 and OCI-1047997 and from the Welch Foundation under Grant No. F-1708.

[*] Corresponding author: dagao2008@gmail.com

[1] J. R. Rodgers and D. Cebon, MRS Bull. **31**, 975 (2006).

[2] K. Rajan, Mater. Today **8**, 38 (2005).

[3] X. He and P. Niyogi, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, B. K. Saul, and B. Scholkopf (MIT Press, Cambridge, MA, 2004).

[4] E. Kokiopoulou and Y. Saad, in *IEEE 5th International Conference on Data Mining (ICDM05), Houston, TX, November 27–30*, edited by J. Han *et al.* (IEEE, New York, 2005), pp. 234–241.

[5] E. Kokiopoulou and Y. Saad, IEEE TPAMI **29**, 2143 (2007).

[6] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction, Information Science and Statistics* (Springer, New York, 2007).

[7] E. Kokiopoulou, J. Chen, and Y. Saad, Numer. Linear Algebra Appl. **18**, 565 (2011).

[8] E. Kokiopoulou and Y. Saad, Pattern Recognition **42**, 2392 (2009).

[9] M. W. Berry, Int. J. Supercomp. Appl. **6**, 13 (1992).

[10] H. Park, P. Howland, and M. Jeon, SIAM J. Matrix Anal. Appl. **25**, 165 (2003).

[11] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).

[12] A. Webb, *Statistical Pattern Recognition*, 2nd ed. (J. Wiley & Sons, Hoboken, NJ, 2002).

[13] T. Hastie, R. Tibshirami, and J. Friedman, *Elements of Statistical Learning, Dataming, Inference, and Prediction, Springer Series in Statistics* (Springer, New York, 2001).

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, *Information Science and Statistics* (Springer, New York, 2006).

[15] J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B **17**, 2453 (1978).

[16] A. N. Bloch and G. Simons, J. Am. Chem. Soc. **94**, 8611 (1972).

[17] J. John and A. N. Bloch, Phys. Rev. Lett. **33**, 1095 (1974).

[18] J. R. Chelikowsky, Phys. Rev. B **26**, 3433 (1982).

[19] P. Villars, J. Less-Common Met. **109**, 93 (1985).

[20] W. Andreoni, G. Galli, and M. Tosi, Phys. Rev. Lett. **55**, 1734 (1985).

[21] W. Andreoni and G. Galli, Phys. Rev. Lett. **58**, 2742 (1987).

[22] J. C. Phillips, *Bonds and Bands in Semiconductors* (Academic Press, New York, 1974).

[23] G. Simons and A. N. Bloch, Phys. Rev. B **7**, 2754 (1973).

[24] W. Andreoni, A. Baldereschi, E. Biémont, and J. C. Phillips, Phys. Rev. B **20**, 4814 (1979).

[25] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[26] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[27] There are four diamond compounds. Two of them are almost in the same location near the point with coordinates $(-2, -0.5)$.

[28] A. N. Tikhonov, Sov. Math. Dokl. **4**, 1036 (1963).

[29] A. N. Tikhonov, Sov. Math. Dokl. **4**, 1624 (1963).

[30] L. Pauling, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed. (Cornell University Press, Ithaca, NY, 1960).