Guiding Scientific Discovery with Explanations using DEMUD

Kiri L. Wagstaff
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91016
kiri.l.wagstaff@jpl.nasa.gov

Nina L. Lanza
Los Alamos National Laboratory
ISR-2, MS D436
Los Alamos, NM 87545
nlanza@lanl.gov

David R. Thompson
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91016
david.r.thompson@jpl.nasa.gov

Thomas G. Dietterich
Oregon State University
1148 Kelley Engineering Center, Corvallis, OR 97331
tgd@eecs.oregonstate.edu

Martha S. Gilmore
Wesleyan University
265 Church St., Middletown, CT 06459
mgilmore@wesleyan.edu

Abstract

In the era of large scientific data sets, there is an urgent need for methods to automatically prioritize data for review. At the same time, for any automated method to be adopted by scientists, it must make decisions that they can understand and trust. In this paper, we propose Discovery through Eigenbasis Modeling of Uninteresting Data (DEMUD), which uses principal components modeling and reconstruction error to prioritize data. DEMUD's major advance is to offer domain-specific *explanations* for its prioritizations. We evaluated DEMUD's ability to quickly identify diverse items of interest and the value of the explanations it provides. We found that DEMUD performs as well or better than existing class discovery methods and provides, uniquely, the first explanations for why those items are of interest. Further, in collaborations with planetary scientists, we found that DEMUD (1) quickly identifies very rare items of scientific value, (2) maintains high diversity in its selections, and (3) provides explanations that greatly improve human classification accuracy.

Introduction

Scientists are increasingly acquiring data sets whose size renders careful examination of each item impractical. Methods for automatically prioritizing data by novelty or scientific interest are vital for making good use of limited analyst time. Equally important, for adoption and for the field of AI, is for such methods to be able to justify their selections with comprehensible explanations.

Our goal is to facilitate scientific discovery, by which we mean "the general process by which scientists discover a new property or learn something new about a natural target or phenomenon," as distinguished from the use of the term in the machine learning field to refer to the induction of scientific laws from data (Langley et al. 1987). Specifically, we focus on the discovery of unusual or unexpected observations within the context of a larger data set.

We propose Discovery through Eigenbasis Modeling of Uninteresting Data (DEMUD) as a strategy for quickly highlighting unusual or interesting items in large data sets. DEMUD uses principal components modeling and reconstruction error to prioritize the data. DEMUD differs from existing anomaly detection methods primarily in its ability to offer accompanying domain-specific *explanations* for why a given item is deemed potentially interesting. These explanations visually depict deviation in the native feature space, and are therefore directly related to physical attributes of the process under study.

We present an illustrative result on a benchmark data set commonly used to evaluate existing methods for "rare category discovery," which is one particular scientific discovery problem. DEMUD performs as well or better than state-of-the-art methods such as SEDER (He and Carbonell 2009) and CLOVER (Huang et al. 2012). Uniquely, DEMUD also offers explanations for its decisions, leading to new and interesting insights even for benchmark data sets.

However, we are unsatisfied with simply conducting benchmark tests. Our ultimate goal is to develop tools that are deemed relevant and useful by the scientific community. Therefore, we have forged collaborations with planetary scientists who have evaluated DEMUD's value and explanatory power in the context of their scientific goals. In experiments with two hyperspectral data sets, we found that DEMUD (1) discovers extremely rare minerals very quickly, (2) maintains a high novelty score, and (3) provides meaningful explanations that greatly increase the accuracy of expert-generated classifications. Scientists have indicated that they particularly value these explanations and that this advance renders the system likely to be adopted for operational use.

Related Work

Our problem formulation is strongly related to anomaly detection, an area of extensive research (Chandola, Banerjee, and Kumar 2009). Common strategies include supervised classification (anomalous vs. normal), rule induction to describe normal items, density-based analysis, clustering, and spectral techniques such as Principal Components Analysis. Common applications include network intrusion, fraud, and disease outbreak detection.

The most relevant work for our purposes is the use of PCA for novelty detection (Hoffmann 2007). PCA-based modeling is generally applied to the entire data set for anomaly

detection (Shyu et al. 2003; Dutta et al. 2007). However, simply ranking all items by an independently computed anomaly score is unlikely to suffice for scientific discovery. It is also important that the results exhibit diversity. That is, subsequent items selected for manual review should take into account those already presented to reduce redundancy and increase the chance of discovering something new.

An emphasis on diversity relates to the problem of rare category detection, in which the goal is to discover an example from every class in the data set as quickly as possible (Pelleg and Moore 2004; He and Carbonell 2007; 2009). These systems iteratively analyze an unlabeled data set to select items that are then labeled by an oracle. Data sets with balanced class distributions can be explored effectively with random selection, but classes with only minority representation require a more complex solution. Such methods include the use of mixture models (Pelleg and Moore 2004) or nearest-neighbor strategies (He and Carbonell 2007), both of which require that the total number of classes be specified in advance. This requirement renders those methods unsuitable for the scientific discovery problem, in which we do not know how many classes are present. SEDER (He and Carbonell 2009), which performs a semiparametric density estimation to discover classes, and CLOVER (Huang et al. 2012), which uses LVD (local variation degree) to improve the computational cost and rate of class discovery, do not require knowledge about the number of classes, but they retain the requirement for a labeling oracle. In contrast, for scientific discovery the user cannot always ascribe a label when presented with a new item. In fact, items that represent a new and previously undiscovered class may be unlabelable without further intensive study. Therefore, we seek a solution that (1) quickly detects novel items and (2) does not require the user to assign category labels.

The final important aspect of this work is its emphasis on providing *human-comprehensible explanations* or justifications for machine-made decisions. Explanation-based learning (Mitchell, Keller, and Kedar-Cabelli 1986) tackles the complementary problem of generating explanations for expert-labeled item classifications by employing a relevant domain theory. Other techniques such as genetic programming can induce simple explanatory rules for classification decisions (Goodacre 2003). To our knowledge, no existing anomaly detection or rare category detection methods have attempted to do this. Strumbelj et al. (2010) proposed the use of per-feature weights to explain per-item classification decisions. Social and natural sciences have long histories of interpreting discriminant models with per-feature loading factors (Betz 1987). These methods are closer in spirit to what DEMUD can provide, but they still depend on the existence of pre-defined classes and labels. DEMUD generates explanations for why each item was selected without such labels.

DEMUD: Discovery via Eigenbasis Modeling of Uninteresting Data

We propose a machine learning solution called Discovery through Eigenbasis Modeling of Uninteresting Data (DEMUD). Here, the term "uninteresting" is a judgment with respect to what should be selected next. It encompasses multiple meanings: data that have already been seen, data that do not fall into a category of interest, or prior knowledge about uninteresting artifacts or behaviors. DEMUD iteratively builds a model of these uninteresting items. This model captures what the user has already seen and should therefore be ignored to increase the chance of selecting a new item of high interest or novelty.

Algorithm 1 DEMUD: Discovery through Eigenbasis Modeling of Uninteresting Data

1: Let $\mathbf{X} \in \mathcal{R}^{(n,d)}$ be the input data set
2: Let $\mathbf{X_U} = \emptyset$ be the set of uninteresting items
3: Let k be the number of principal components used to model $\mathbf{X_U}$
4: Let $\mathbf{U}, \mu = \text{SVD}(\mathbf{X}, k)$ be the initial model of $\mathbf{X_U}$ and the data mean μ
5: **while** patience remains and $\mathbf{X} \neq \emptyset$ **do**
6: Compute reconstructions $\hat{\mathbf{x}} = \mathbf{U}\mathbf{U}^T(\mathbf{x} - \mu) + \mu$ for all $\mathbf{x} \in \mathbf{X}$
7: Update scores $S_x = R(\mathbf{x})$ with Eqn. 2 for $\mathbf{x} \in \mathbf{X}$
8: Select $\text{argmax}_{\mathbf{x}' \in \mathbf{X}} S_x$
9: Create per-feature explanations $e_j = \mathbf{x}'_j - \hat{\mathbf{x}}'_j$ for $j = 1 \ldots d$
10: $\mathbf{X} = \mathbf{X} \setminus \{\mathbf{x}'\}$
11: $\mathbf{X_U} = \mathbf{X_U} \bigcup \{\mathbf{x}'\}$
12: **if** $|\mathbf{X_U}| == 1$ **then**
13: Let $\mathbf{U}, \mu$ = SVD($\mathbf{X_U}, k$)
14: **else**
15: Update $\mathbf{U}, \mu$ = incremSVD($\mathbf{U}, \mathbf{x}', k$)
16: **end if**
17: **end while**

There are several possible ways to model the uninteresting items. For scalability to large data sets, we selected a linear method that can be efficiently and incrementally updated. We compute a low-dimensional eigenbasis representation of the uninteresting items via Singular Value Decomposition (SVD): ${\mathbf{X_U}}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. We retain the top k vectors in $\mathbf{U}$, ranked by the magnitude of the corresponding singular values, then use this model to rank the remaining items by their reconstruction error. Items with high error are those that are poorly modeled by $\mathbf{U}$ and therefore have the highest potential to be novel.

DEMUD is an iterative strategy (see Algorithm 1). At each iteration, the items in $\mathbf{X}$ have not (yet) participated in the construction of $\mathbf{U}$, since $\mathbf{X}$ and $\mathbf{X_U}$ are disjoint. Therefore we reconstruct each $\mathbf{x}$ as $\hat{\mathbf{x}}$ by projecting $\mathbf{x}$ onto $\mathbf{U}$ and then back into the original feature space. The score for $\mathbf{x}$ is the reconstruction error between $\mathbf{x}$ and $\hat{\mathbf{x}}$ (lines 6–7):

$$R(\mathbf{x}) = ||\mathbf{x} - \hat{\mathbf{x}}||_2 \quad (1)$$

$$= ||\mathbf{x} - (\mathbf{U}\mathbf{U}^T(\mathbf{x} - \mu) + \mu)||_2, \quad (2)$$

where μ is the mean of all previously seen $\mathbf{x} \in \mathbf{X_U}$. The first iteration, in which $\mathbf{X_U}$ is empty, uses the full data set mean for μ and $\mathbf{U}$ from the full data set decomposition (line 4). The top-scoring observation $\mathbf{x}$ is selected (line 8).

Next, DEMUD creates per-feature explanations that are

the residual values, i.e., the difference between the true value and the reconstructed one, which is the information that the model could not explain (line 9). These are discussed further in the next section. Then $\mathbf{x}$ is removed from $\mathbf{X}$ (line 10) and added to the set of uninteresting items (line 11).

DEMUD's model $\mathbf{U}$ is initially computed from the whole data set (line 4) to provide a default ranking of the data. The first iteration performs an SVD on the single item in $\mathbf{X}_\mathbf{U}$ (line 13), and subsequent iterations update this $\mathbf{U}$ with the new $\mathbf{x}'$ (line 15). DEMUD uses a fast, incremental SVD technique (Lim et al. 2005) that improves over the popular R-SVD algorithm (Golub and Van Loan 1996) by also tracking changes in the sample mean μ induced by the inclusion of $\mathbf{x}'$:

$$\mu' = \frac{n}{n+1}\mu + \frac{1}{n+1}\mathbf{x}', \tag{3}$$

where n is the number of items that contributed to the existing $\mathbf{U}$. It then passes $\mathbf{U}$ and an augmented matrix $\left[\mathbf{x}' - \mu' | \sqrt{\frac{n}{n+1}}(\mu - \mu')\right]$ to R-SVD to obtain the new $\mathbf{U}$.

DEMUD with Explanations

The explanations generated by DEMUD (line 9 in Algorithm 1) express the degree to which each feature value deviated from the model's expectations. This is computed as the difference between the observed value and the value predicted by the model via reconstruction. As we will see in the Results section, these quantitative explanations can provide powerful interpretative insights when combined with domain knowledge about the features.

DEMUD's explanations differ from methods for feature selection, which identify features that are relevant for all items in the data set. DEMUD's explanations are not only item-specific, but also context-specific. Since the model $\mathbf{U}$ is updated after each item is selected, the explanations indicate why each feature value is anomalous *with respect to what has been seen before*. Those judgments change with each iteration, tracking what the human reviewer has already seen and (presumably) learned.

Experimental Results

We conducted experiments to evaluate (1) DEMUD's ability to discover novel items within a data set and (2) the utility of the explanations generated by the system.

Benchmark class discovery

As previously discussed, class discovery is one kind of scientific discovery that we aim to facilitate. Evaluating DEMUD on this task also allows us to compare directly with existing techniques that depend on the definition of discrete classes, even though DEMUD does not.

The glass data set can be found in the UCI repository (Frank and Asuncion 2010). It consists of 214 glass fragments that are described by 9 features: their refractive index (RI) and 8 compositional features (Na, Mg, Al, Si, K, Ca, Ba, and Fe content). There are six types of glass in the data set: building windows that are float processed, building windows that were not float processed, vehicle windows that were float processed, containers, tableware, and headlamps. This has been one of the data sets of choice for use by researchers working on rare category detection methods.

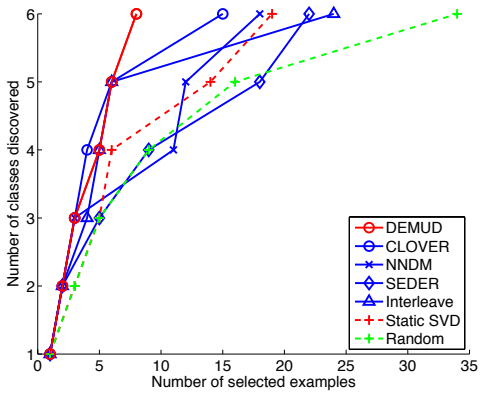


Figure 1: Class discovery rate for the glass data set ($n = 214, d = 9$). Methods that learn iteratively (solid lines) tend to perform better than static methods (dashed). CLOVER is the best-performing previous method but requires a labeling oracle. DEMUD discovers all 6 classes in only 8 selections without an oracle or knowledge of the number of classes.

Class discovery. Figure 1 shows the empirical class discovery rate for the glass data set. Results for CLOVER and random sampling were obtained from Huang et al. (2012); results for SEDER, NNDM, and Interleave were obtained from He and Carbonell (2009). The figure shows results for DEMUD ($k = 2$) and for a static baseline strategy that ranks all items by their reconstruction error using the full data set SVD ($k = 2$). DEMUD differs in that it iteratively updates its model to select items that are different from those already selected. CLOVER achieved the previous best result in discovering all classes with 15 items selected; DEMUD required only 8.

Although CLOVER and SEDER are "prior-free" in that they do not require advance knowledge of how many classes there are, they do require feedback for each selection before proceeding to the next. DEMUD (and the static SVD) require no feedback to continue exploring the data set. Of course, an oracle or reviewer is required to assess whether a new class has been discovered, but this can be done after DEMUD processes the whole data set, rather than requiring that the oracle be available in the processing loop. Further, the oracle's time can be employed to assess only the top-ranked items rather than required for the whole data set.

Glass explanations. In addition to discovering classes, DEMUD provides an explanation for why each item was selected. Table 1 shows the explanations associated with each new class discovered, in terms of the item's residuals (the difference between the observed and reconstructed values). We highlight residuals greater than 1.00 (percent composition) for emphasis (positive in green, negative in red). For example, the first example of container glass is strongly en-

Table 1: DEMUD (using $k = 2$) explanations for each class discovered in the glass data set, expressed as residuals in original units (percent composition). Positive (negative) values are higher (lower) than expected; residuals with absolute value greater than 1.00 are highlighted. The first discovery of an example of each of the 6 classes is shown. All 6 classes were discovered after 8 selections from the data set of 214 items, which is the best published result.

Selection	Class (proportion)	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	container (6%)	−0.001	−1.60	−0.86	+0.79	−2.80	+5.40	−0.24	−0.88	−0.01
2	building window, non-float (36%)	0.000	−0.72	0.00	−2.00	−0.32	−6.10	+9.20	0.00	+0.24
3	tableware (4%)	+0.005	+4.60	0.00	−2.10	+5.00	−4.50	−2.90	0.00	−0.07
5	headlamp (14%)	−0.002	−2.80	−0.56	−0.41	+3.00	+1.30	−0.05	−0.46	−0.06
6	building window, float (33%)	+0.003	−0.28	+4.00	−0.50	−0.80	−1.80	−0.32	−0.37	−0.05
8	vehicle window, float (8%)	+0.002	+0.43	+2.90	−0.43	−1.20	−0.93	−0.02	−0.49	−0.07

riched in K but depleted in Na and Si, *with respect to the overall data set*. The first example of building window glass (non-float treated) is strongly enriched in Ca and much lower in K and Al, *with respect to the container glass* already seen. Item 3 (first example of tableware glass) is enriched in Na and Si but depleted in Al, K, and Ca *with respect to the first two items*. Each item's annotations explain why it was chosen and can aid in interpretation.

To our knowledge, this is the first attempt to explain the contents of the glass data set. The 1987 paper that introduced the data set motivated its study from the perspective of forensic science: the ability to classify the type of a glass fragment could help solve crimes (Evett and Spiehler 1987). We have not seen any evidence of a machine learning system that has actually been employed by forensic science to this purpose, nor even any content-focused discussion of the data set. Papers that use the data set to evaluate class discovery methods do not report the order in which the classes are discovered, so we do not know whether the order in which DEMUD discovered them is typical. We welcome further comparisons.

Scientific discovery

Although DEMUD can be useful for class discovery, it was designed for the more challenging and less constrained problem of scientific discovery, in which no labels are available to guide the exploration of the data set. We simulated the discovery process by identifying a subset of items with high scientific interest, then assessed how quickly DEMUD and other strategies found those items. Finally, we evaluated the utility of DEMUD's explanations by measuring their influence on human classification performance.

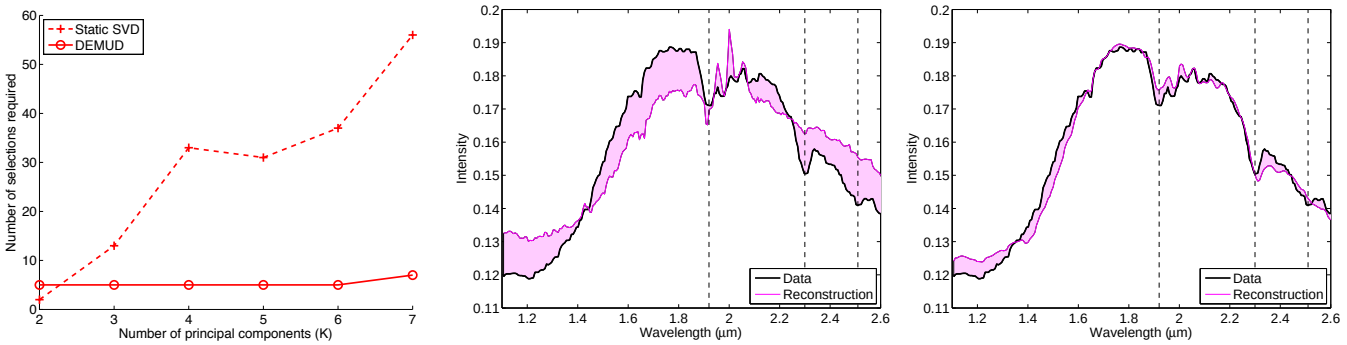
Baseline: Outlier selection. In addition to random selection as a default baseline, we also compared DEMUD to a simple outlier detection strategy. This method ranks all items in descending order of their Euclidean distance from the data set mean. Like the static SVD, it is done once rather than updating iteratively like DEMUD. More sophisticated strategies exist, but none provide explanations, as noted earlier.

(1) CRISM hyperspectral observations of Mars. Our first application is the exploration of hyperspectral data collected from planetary orbit. The Compact Reconnaissance Imaging Spectrometer (CRISM) onboard the Mars Reconnaissance Orbiter spacecraft has collected hundreds of very large hyperspectral images of the surface (Murchie et al. 2007). These spectra reveal a wide range of physical phenomena including atmospheric effects, surface thermal emissive properties, and rich mineralogical insights. Working with a domain expert, we chose to study CRISM scene frt00003e12 of Nili Fossae[1] which contains small isolated magnesite ($MgCO_3$) deposits, a carbonate that forms in the presence of water (Ehlmann et al. 2009). To reduce noise and data set size, without losing important details, we first performed a superpixel segmentation to split the image into several thousand homogeneous segments, each represented by its mean spectrum (Thompson et al. 2010), and then used a median filter to remove shot noise. The resulting large data set consists of 26,500 superpixels with 230 features covering the range 1.1 to 2.6 μm (near infrared). The 18 magnesite examples constitute only 0.06% of the data.

Magnesite discovery. We evaluated DEMUD and other methods in terms of the number of items selected to achieve the first discovery of magnesite (lower is better). Figure 2a shows the magnesite discovery results for k values from 2 to 7 (with $k = 1$, all methods took more than 500 selections to find the magnesite). In general, DEMUD requires just 5 selections to find the magnesite, while the static SVD typically takes far more. An exception is at $k = 2$, where the static SVD found the magnesite with only 2 selections. As k increases, both methods are able to model more data complexity. However, for the static SVD, this rendered the magnesite more difficult to find. In contrast, DEMUD robustly and efficiently detected the magnesite for all k values shown. DEMUD does show a gradual increase in the number of selections needed with larger k, but it is far less sensitive than the static SVD; we truncated the plot at $k = 7$ because at $k = 8$ the static SVD required 209 selections (DEMUD required only 9). Both methods are superior to random and outlier-based selection, which do not vary with k; they required 1472 and 7227 selections respectively.

Magnesite explanations. It is easy to "discover" that a spectrum is magnesite when it comes with a label. In a real setting, however, the scientist must examine each spectrum to determine what it might contain. Our goal is to accelerate

[1]Data available at http://imbue.jpl.nasa.gov/ .



(a) Effort required to discover magnesite. (b) DEMUD's magnesite explanations. (c) Static SVD's magnesite explanations.

Figure 2: Magnesite discovery in CRISM data ($n = 26500, d = 230$). Panel (a) plots the number of selections required to find the first magnesite sample (lower is better). Explanations for the first discovered magnesite are shown for (b) DEMUD and (c) static SVD (both using $k = 3$). Selected absorption bands are plotted with dashed lines for aid in interpretation. DEMUD shows large residuals at 2.3 and 2.5 μm, which are diagnostic of magnesite, while the uninformative band at 1.9 μm has a small residual. The static SVD shows small residuals at all bands and therefore provides no salient explanation for this item.

this process by providing explanations that point the scientist to diagnostic features of interest.

Figure 2 shows the residual-based explanations provided by (b) DEMUD and (c) the static SVD. Selected absorption bands are indicated with dashed lines to aid in interpretation. The band at 1.9 μm is present throughout the data set and therefore not diagnostic for this sample. DEMUD, at item 5, had already learned and incorporated this into its model after only four previous selections, so it has a very small residual. The bands at 2.3 and 2.5 μm, in contrast, have large residuals and are diagnostic of magnesite specifically. These residuals provide exactly the right domain-specific answer to, "Why is this spectrum interesting?" Large residuals are also seen between 1.5 and 1.8 μm, but these are due to overall brightness rather than to absorption bands that convey compositional information. In contrast, the static SVD, shows small residuals at all bands and therefore provides no salient explanation for this item (its 13th selection).

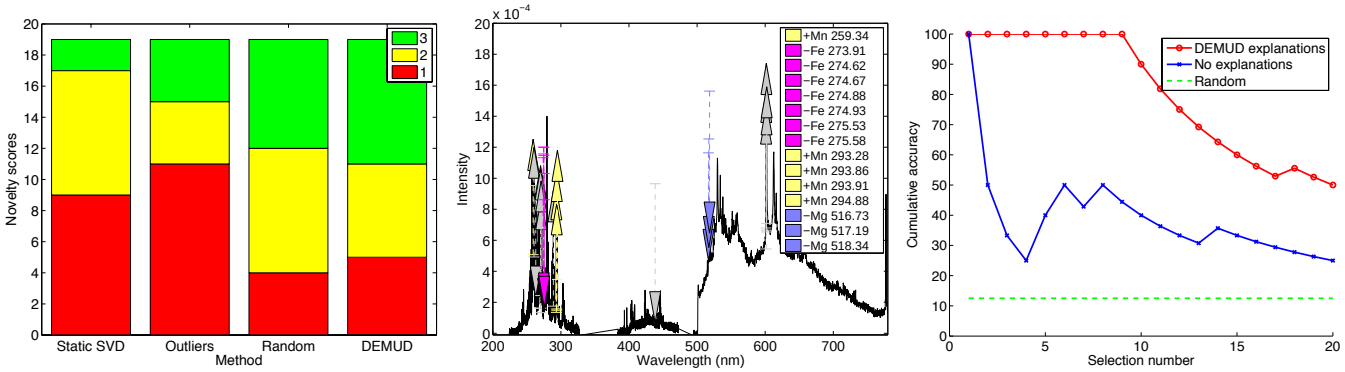
Naturally, if it were known in advance that magnesite was going to appear in this data set, one could (and scientists do) compute the similarity of each sample to a known magnesite spectrum. But in practice, such analyses are limited to a finite set of likely minerals, and they leave open the question of what is not detected because it was not anticipated. The magnesite example simulates the more general setting in which the contents of a large data set are not known in advance and finding unanticipated items can lead to new scientific discoveries.

(2) ChemCam Martian point spectra. The ChemCam instrument on the Mars Science Laboratory rover uses a Laser-Induced Breakdown Spectrometer (LIBS) to obtain spectroscopic observations, using 6144 bands from 224 to 932 nm, of targets up to 7 meters away (Wiens, Maurice, and the ChemCam team 2011). In contrast to CRISM's reflectance spectra, ChemCam acquires *emission* spectra from targets stimulated by its laser. These spectra can indicate the presence of individual elements. We applied DEMUD to a data set collected using ChemCam calibration materials on Earth (Lanza et al. 2010). It contains 110 spectra consisting of eight sample types. This data set provides a complementary challenge to the CRISM data set because it combines extremely high dimensionality with fewer distinct samples.

Selection novelty. Rather than searching for a sample of interest, we used this data set to assess the diversity and novelty of DEMUD's selections. We collected the first 20 spectra selected by DEMUD, the static SVD, the outlier baseline, and random selection. We presented the selections to an expert from the ChemCam science team who was asked to rate each item subjectively on a score from 1 to 3, where 1 means "redundant" with an earlier selection, 2 means "some novel features", and 3 means "novel, possibly a new mineral type." The sets selected by the four methods were presented in an arbitrary, unlabeled order so that the scientist was unaware of which method was used to generate each set. To remove any obvious indicators, the first selection in each list was fixed to be the item with largest reconstruction error with respect to the whole data set; scores are reported for items 2 through 20. We also encouraged the expert to take breaks between sets to mentally "reset" to a blank slate.

Figure 3a shows the distribution of novelty scores achieved by each method. DEMUD achieved the largest number of "3" scores, closely followed by random selection. The static SVD and outlier-based results were dominated by low scores. Further, DEMUD achieved high novelty scores while also providing explanations for its decisions, something the random selection method cannot do.

ChemCam explanations. The emission spectra observed by ChemCam provide elemental abundance information based on individual bands, so DEMUD's explanations can be further tailored to automatically interpret large residuals. For example, Figure 3b shows item 9 chosen by DEMUD, its first discovery of rhodochrosite ($MnCO_3$). The top 5% of



(a) Distribution of novelty scores (b) Rhodochrosite ($MnCO_3$) with explanations (c) Expert classification performance

Figure 3: DEMUD results on ChemCam data ($n = 110, d = 6143$): (a) novelty scores according to a member of the ChemCam science team (higher is better); (b) sample output with explanations; (c) improvements in human classification performance obtained when using DEMUD's explanations.

residuals (by magnitude) are indicated with arrows that point from the reconstructed (expected) value to the actual observed value; each is associated with an element, if known. The explanations indicate that this spectrum has higher than expected evidence for Mn and lower than expected Fe and Mg. Earlier selections included siderite ($FeCO_3$) and olivine (high in Mg). Also note that DEMUD doesn't simply annotate all large peaks; in fact, many of the bands highlighted have small values. Unannotated peaks are common features that DEMUD has learned to ignore. While somewhat cryptic to the non-geochemist, these annotations zero in on vital diagnostic clues for the expert.

Finally, we measured the utility of DEMUD's explanations in terms of their impact on the expert's ability to classify the data (a challenging task without supplemental information). We asked the ChemCam scientist to manually classify DEMUD's first 20 selections into eight categories: andesite, basalt, calcite, dolomite, limestone, olivine, rhodochrosite, and siderite. The classifications were done first without, and then with, DEMUD's explanations. We found that overall accuracy doubled from 25% to 50% using DEMUD's explanations; random performance was 13%. Figure 3c shows that the benefits were strongest for the earliest selections: cumulative accuracy increased from 40% to 90% for the first 10 selections. Further study is needed to explain this effect as due to earlier samples being easier to classify, earlier explanations highlighting larger differences between samples, data classification fatigue, or some combination of these factors. Regardless, this experiment provides concrete evidence that DEMUD's explanations are meaningful, appropriate, and useful.

Conclusions

DEMUD is a method for discovering novel observations in large data sets that also provides an explanation for why each item is selected. It does so using an efficient, incremental SVD method that progressively learns a model of what has already been selected so that items with high novelty (measured by reconstruction error) can be selected next. We found that DEMUD performs well in (1) discovering new classes, (2) finding extremely rare samples of interest (e.g., CRISM magnesite), (3) selecting items with high novelty, and (4) providing salient, useful explanations for each one. We demonstrated a concrete benefit to scientists in observing that DEMUD's explanations led to a large improvement in the accuracy of expert classification of ChemCam spectra.

DEMUD can make better use of limited human review time by focusing attention on the most unusual items first. It can provide a complement to other strategies for analyzing the data, such as (supervised) searches for known targets of interest. Further, the concept of "uninteresting" items is broad enough that it can be used to express prior knowledge in the form of observations that are already known, so that DEMUD will seek very different ones. We will explore this angle in future work, as well as the use of a robust incremental SVD to reduce sensitivity to noise (Li 2004).

The explanations provided by DEMUD are the primary novel contribution of this work. Given the choice, we find that scientists strongly prefer results accompanied by explanations to those that come from a mute black box. DEMUD's domain-specific explanations greatly increase data interpretability and therefore the likelihood of the system's adoption by scientists and users outside the machine learning community.

Acknowledgements

We wish to thank Roger Wiens, Diana Blaney, and Sam Clegg for their assistance with the ChemCam data and the Planetary Data System (PDS) for providing the CRISM data. This work was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

References

Betz, N. E. 1987. Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology* 34(4):393.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys* 41(3):15:1–15:58.

Dutta, H.; Giannella, C.; Borne, K.; and Kargupta, H. 2007. Distributed top-k outlier detection in astronomy catalogs using the DEMAC system. In *Proceedings of the SIAM International Conference on Data Mining*.

Ehlmann, B. L.; Mustard, J. F.; Swayze, G. A.; Clark, R. N.; Bishop, J. L.; Poulet, F.; Marais, D. J. D.; Roach, L. H.; Milliken, R. E.; Wray, J. J.; Barnouin-Jha, O.; and Murchie, S. L. 2009. Identification of hydrated silicate minerals on Mars using MRO-CRISM: Geologic context near Nili Fossae and implications for aqueous alteration. *Journal of Geophysical Research* 114(E00D08). doi:10.1029/2009JE003339.

Evett, I. W., and Spiehler, E. J. 1987. Rule induction in forensic science. In *KBS in Goverment*, 107–118. Online Publications.

Frank, A., and Asuncion, A. 2010. UCI machine learning repository. http://archive.ics.uci.edu/ml.

Golub, G. H., and Van Loan, C. F. 1996. *Matrix Computations*. The Johns Hopkins University Press.

Goodacre, R. 2003. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrational Spectroscopy* 32(2):33–45.

He, J., and Carbonell, J. 2007. Nearest-neighbor-based active learning for rare category detection. In *In Advances in Neural Information Processing Systems 21*, 633–640. MIT Press.

He, J., and Carbonell, J. 2009. Prior-free rare category detection. In *Proceedings of the SIAM International Conference on Data Mining*, 155–163.

Hoffmann, H. 2007. Kernel PCA for novelty detection. *Pattern Recognition* 40(3):863–874.

Huang, H.; He, Q.; Chiew, K.; Qian, F.; and Ma, L. 2012. CLOVER: a faster prior-free approach to rare-category detection. *Knowledge and Information Systems*. doi:10.1007/s10115-012-0530-9.

Langley, P. W.; Simon, H. A.; Bradshaw, G. L.; and Zytkow, J. M. 1987. *Scientific Discovery: Computational Explorations of the Creative Process*. MIT Press.

Lanza, N. L.; Wiens, R. C.; Clegg, S. M.; Ollila, A. M.; Humphries, S. D.; Newsom, H. E.; and Barefield, J. E. 2010. Calibrating the ChemCam laser-induced breakdown spectroscopy instrument for carbonate minerals on Mars. *Applied Optics* 49(13):C211–C217.

Li, Y. 2004. On incremental and robust subspace learning. *Pattern Recognition* 37:1509–1518.

Lim, J.; Ross, D.; Lin, R.-S.; and Yang, M.-H. 2005. Incremental learning for visual tracking. In *Advances in Neural Information Processing Systems 17*, 793–800. MIT Press.

Mitchell, T. M.; Keller, R. M.; and Kedar-Cabelli, S. T. 1986. Explanation-based generalization: A unifying view. *Machine Learning* 1(1):47–80.

Murchie, S.; Arvidson, R.; Bedini, P.; Beisser, K.; Bibring, J.; Bishop, J.; Boldt, J.; Cavender, P.; Choo, T.; Clancy, R.; et al. 2007. Compact reconnaissance imaging spectrometer for Mars (CRISM) on Mars Reconnaissance Orbiter (MRO). *Journal of Geophysical Research* 112(10.1029).

Pelleg, D., and Moore, A. 2004. Active learning for anomaly and rare-category detection. In *In Advances in Neural Information Processing Systems 18*, 1073–1080. MIT Press.

Shyu, M.-L.; Chen, S.-C.; Sarinnapakorn, K.; and Chang, L. 2003. A novel anomaly detection scheme based on principal component classifier. In *Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop*, 172–179.

Strumbelj, E.; Bosnic, Z.; Kononenko, I.; Zakotnik, B.; and Kuhar, C. G. 2010. Explanation and reliability of prediction models: the case of breast cancer recurrence. *Knowledge and Information Systems* 24:305–324.

Thompson, D.; Mandrake, L.; Gilmore, M.; and Castaño, R. 2010. Superpixel endmember detection. *IEEE Transactions on Geoscience and Remote Sensing* 48(11):4023–4033.

Wiens, R. C.; Maurice, S.; and the ChemCam team. 2011. The ChemCam instrument suite on the Mars Science Laboratory rover Curiosity: Remote sensing by laser-induced plasmas. *Geochemical News* 145.