On-board Analysis of Uncalibrated Data for a Spacecraft at Mars

Rebecca Castano
rebecca.castano@jpl.nasa.gov

Kiri L. Wagstaff
kiri.wagstaff@jpl.nasa.gov

Steve Chien
steve.chien@jpl.nasa.gov

Timothy M. Stough
timothy.stough@jpl.nasa.gov

Benyang Tang
benyang.tang@jpl.nasa.gov

Jet Propulsion Laboratory
California Inst. of Technology
4800 Oak Grove Drive,
Pasadena, CA 91109

ABSTRACT

Analyzing data on-board a spacecraft as it is collected enables several advanced spacecraft capabilities, such as prioritizing observations to make the best use of limited bandwidth and reacting to dynamic events as they happen. In this paper, we describe how we addressed the unique challenges associated with on-board mining of data as it is collected: uncalibrated data, noisy observations, and severe limitations on computational and memory resources. The goal of this effort, which falls into the emerging application area of spacecraft-based data mining, was to study three specific science phenomena on Mars. Following previous work that used a linear support vector machine (SVM) on-board the Earth Observing 1 spacecraft, we developed three data mining techniques for use on-board the Mars Odyssey spacecraft. These methods range from simple thresholding to state-of-the-art reduced-set SVM technology. We tested these algorithms on archived data in a flight software testbed. We also describe a significant, serendipitous science discovery of this data mining effort: the confirmation of a water ice annulus around the north polar cap of Mars. Finally, we conclude with a discussion on lessons learned in developing algorithms for use on-board a spacecraft.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

Keywords

on-board data mining, real-time data analysis, resource-constrained computing, support vector machines, lessons learned

1. INTRODUCTION

There are a number of resource-constrained environment application domains in which it may be desired to perform data mining. Examples include embedded mobile devices, sensor webs, and on-board settings such as spacecraft or unpiloted air vehicles. Typically, these environments have limited CPU and RAM resources. In many cases, the data available is uncalibrated. An emerging paradigm in this category, addressed in this paper, is the mining of scientific data on-board a spacecraft. A primary purpose of scientific planetary spacecraft is to collect measurements of physical values that provide information to gain an understanding of the world (e.g. space environment, planetary surface or atmosphere, etc.) Historically, data is collected by prespecifying all possible parameters including when and where to acquire measurements. Collected data is transmitted to the ground where it is calibrated and then analyzed by domain experts such as scientists. There is no opportunity to adjust what is acquired or transmitted based on the contents of the data itself.

Recently, we posited that significantly more science could be accomplished by a mission or spacecraft if data can be analyzed on-board, and we demonstrated the principles through the use of a linear support vector machine (SVM) on-board the Earth orbiting EO-1 spacecraft [4, 3]. There are three major reasons that on-board science data mining is desirable. First, it can enable detection and rapid reaction to dynamic events. Second, on-board data analysis can be used to prioritize the data that is collected. Finally, that analysis can produce additional science products (e.g., data summaries or dynamic event detections) that consume little bandwidth but provide key insights into the data. We focus on the latter two cases, which both enable more effective use of limited bandwidth.

A common property of remote missions is that the spacecraft or instrument is capable of collecting more data than can be accommodated by the downlink volume available. In this circumstance, an opportunity exists to collect additional data and select only the most interesting for transmission. Success of such an operational mode relies on being able to identify, on-board, data that is more interesting than what

would be collected with the traditional method of preselecting exactly what data to collect and when. This requires a specification of what constitutes "interesting" data, which must be translated into an algorithm that can operate in the computational environment of the spacecraft on the data in the form it is available (i.e., uncalibrated, noisy).

In this paper, we present three data mining techniques for use on-board the Mars Odyssey spacecraft, ranging from simple thresholding to state-of-the-art reduced-set SVM technology. Each of these algorithms has been thoroughly tested on archived data in a flight software testbed. We describe the algorithms and test results, including a significant, serendipitous scientific finding that resulted from the data mining effort: the confirmation of a water ice annulus around the north polar cap of Mars. We also present lessons learned in developing successful on-board science data analysis applications.

2. APPLICATION DOMAIN: MARS DATA ANALYSIS

This paper focuses on techniques that have been developed for three science investigations relevant to data that is collected in Mars orbit by the Mars Odyssey spacecraft. Odyssey was launched in 2001 and has been mapping the surface of Mars for more than five years. We focus specifically on observations made by the Thermal EMission Imaging System (THEMIS), a multi-wavelength camera on Odyssey [5]. THEMIS combines a 5-wavelength visible imaging system (0.425–0.860 μm) with a 10-band infra-red (IR) imaging system (6.78–14.88 μm). The resolution of the visible imager is 18 meters per pixel while the resolution of the IR imager is 100 meters per pixel. In this work, we analyze the THEMIS IR images. Each image is 320 pixels (32 km) wide and a variable number (3600 to 14352) of pixels long, divided into 256-line "framelets".

A motivation for selecting the THEMIS instrument is that it does have the capability of collecting more data than bandwidth limitations will permit to be downloaded to Earth. By analyzing data collected at the full capability of the instrument, up to an order of magnitude more area may be monitored for rare science features. There are two important challenges associated with analyzing THEMIS data. The first is that data is not calibrated on-board the spacecraft, so any analysis performed must be robust to significant noise. Second, the THEMIS camera experiences significant "drift", in which the camera's response function is altered due to temperature changes it experiences during its orbit. As a result, the values it records over the course of a single image can gradually increase or decrease even when there is no actual change being observed. This phenomenon is more pronounced for longer images.

We focus on an operational scenario in which we wish to make the best use of limited bandwidth. Each of the three methods that we have developed makes a determination about whether an event of interest is contained in a given THEMIS image. If there is a positive detection, there are several options for how to proceed. In order of lowest to highest bandwidth use, they are:

1. Transmit a brief summary of the detection, such as the latitude and longitude or time-on-orbit when the detection was made.

2. Transmit a subset of the image that covers the region that caused the positive detection.

3. Transmit the entire image when any detection is made.

Depending on the current bandwidth available, mission operators can select the appropriate mode for operation. Option 1 requires the least bandwidth can can be used once operational accuracy has been validated. Option 2 requires less bandwidth than option 3 but higher complexity, since it requires the ability to crop or subset an image after it is collected; this capability was not originally designed into the Mars Odyssey software and would need to be added.

The specific science goals that we seek to support through on-board data mining are: thermal anomaly detection, polar cap tracking, and aerosol opacity estimation. We discuss each application, the algorithmic solution, and provide test results for each method.

3. THERMAL ANOMALY DETECTION

The first algorithm was developed to identify thermal anomalies on the surface of the planet. This feature was selected due to the profound scientific significance of detecting such an anomaly. It is not definitively known whether Mars is currently thermally active. Obtaining proof of current thermal activity would have major scientific implications. A thermal anomaly is defined as a region where the surface temperature is significantly warmer or colder than expected, given its location on the planet, the season, and local topography. A warm thermal anomaly could indicate the presence of subsurface hydrothermal activity, which would have immediate implications for the search for life. No such regions are yet known to exist on Mars, and they are likely to be small and rare, if present at all. Nevertheless, it is difficult to imagine a more significant discovery about the Martian surface, short of detecting large amounts of water or life itself. Other thermal anomalies include active lava flows, frost at low latitudes, and very fresh impact craters. THEMIS, with its high spatial resolution and thermal sensitivity, is an excellent instrument for searching for thermal anomalies.

3.1 Thermal Anomaly Detection: Algorithm

Our approach detects thermal anomalies by searching for pixels that exceed or drop below a specified temperature threshold. To estimate the temperature, we use a single wavelength band: THEMIS band 9 (12.57 μm), where the instrument has the greatest signal to noise ratio and is most sensitive to surface temperatures. We use an approximate conversion from temperature to a specific DN (raw pixel) value. The particular threshold used can vary depending on the type of thermal anomaly, the time of day (nighttime vs. daytime), the latitude, and the season. Scientists will be able to specify these parameters from the ground, however a single threshold is used per image. As there may be data artifacts or noisy pixels within an image, we apply a post-processing step to minimize false alarms. If there are more than a specified number of pixels that are above the threshold, the image is not flagged as containing a thermal anomaly. This post-processing is employing the domain

knowledge that any thermal anomaly discovered should be very localized.

3.2 Thermal Anomaly Detection: Results

We evaluated this algorithm on 14,856 archived THEMIS images with the goal of detecting hot thermal anomalies. We analyzed nighttime images that were collected between 60 degrees north and 60 degrees south of the equator, with a threshold of 240 K. The thermal anomaly detection method signaled a detection for 143 of the 14,856 images. As no thermal anomalies are yet known to exist on Mars, these detections can be considered to be false alarms. However, the domain scientist expressed an interest in manually examining these images and found them to be interesting from a geological perspective. Thus, the global thresholding reduced the number of image candidates requiring manual analysis by a factor of 100. Operationally, the false alarm rate ($< 1\%$) was deemed acceptable by the THEMIS scientists. Since no true positive examples are available, we also conducted a set of tests to confirm detection with synthetically introduced thermal anomalies. All such synthetic positives were correctly detected by the algorithm.

4.   POLAR CAP EDGE DETECTION

The second algorithm was developed to identify polar cap edges. Like the Earth, Mars experiences significant seasonal weather patterns. One result of these changes is the presence of CO_2 ice caps at both poles that advance and recede seasonally. The seasonal cycling of CO_2 from the atmosphere to the polar caps (condensation) and back to the atmosphere (sublimation) significantly alters the distribution of mass on the planet. This effect is large enough that it is possible, even from Earth, to observe the resulting oscillation of the center of gravity of Mars [10]. Scientists are interested in tracking the motion of the polar caps over time so that we can better understand the processes at the north and south poles as well as any interannual changes in polar cap behavior. Since the polar caps stand out as distinctly colder than the rest of Mars, an ideal way to track them is to use an camera in Mars orbit. THEMIS has yielded the best IR observations of the polar cap in terms of spatial resolution (100 m per pixel), exceeding that of the Thermal Emission Spectrometer (TES) (3 km per pixel) on the Mars Global Surveyor spacecraft and the Visible and Infrared Mapping Spectrometer (OMEGA) (300 m to several km per pixel) on the Mars Express spacecraft.

Since the exact location of the edge of the polar cap is not always known ahead of time, not every polar image successfully captures it. For example, a recent north pole imaging campaign resulted in a data set that is about one-third composed of images containing the polar cap edge. With the ability to automatically prioritize images that contain the cap edge, we can increase the number of detections that are transmitted to Earth without increasing, or even by decreasing, the amount of bandwidth required.

As with thermal anomaly detection, we use THEMIS band 9 data to obtain the surface temperate. Some details of the polar cap edge detection algorithm were previously reported at the i-SAIRAS conference [14]. We present new results obtained when porting this algorithm to a flight software testbed. We also discuss the serendipitous discovery of the water ice annulus south of the north polar cap on Mars.

4.1   Polar Cap Edge Detection: Algorithm

Due to Mars Odyssey's orbit, images of the north polar region are collected from north to south on the daylit side of the planet. An image may partially contain the polar cap, it may not contain the cap at all, or it may contain only the cap. Our algorithm exploits the fact that the defrosted terrain surrounding the polar cap is significantly warmer than the CO_2 frost and ice. Therefore, we can detect whether a given image contains the edge of the polar cap by determining whether its temperature histogram is bimodal. If it contains two peaks (at reasonable temperatures), then it is likely to contain the edge of the polar cap; if not, it is likely to cover only the polar cap or only the defrosted terrain. For bimodal images, once we identify a threshold between the peaks, we can identify the location in the image where the transition from CO_2 ice to defrosted terrain (the edge of the cap) occurs.

The true condensation temperature of CO_2 is known *a priori*. If the temperature data available on-board the spacecraft were fully calibrated, we could specify the expected separation between "CO_2 ice" and "defrosted terrain". However, since we are working with uncalibrated data as it is collected, we must instead adaptively determine what threshold to use.

Step 1: Calibration [Optional]. Our method of characterizing the shape of the temperature histogram does not rely on absolute temperatures and, in fact, can be applied without any calibration. However, we gain a slight improvement in precision by performing a fast approximate calibration to help select the most appropriate threshold. We pseudo-calibrate [1] each pixel i in the image by converting the raw digital number, DN_i, to a temperature, T_i (Kelvin):

$$x = (DN_i - o \times g) \times g/16$$
$$T_i = 101.85 \times \ln(x) - 223.3,$$

where o and g are the instrument offset and gain parameters, provided in the header of the data file.

Step 2: Temperature Histogram. We construct a histogram of all of the temperature values in the image. Each histogram bin is 2 K wide, and the histogram ranges from 130 to 270 K. Figure 1(a) shows the histogram for image I09779015. The larger mode corresponds to the cold areas covered by frozen CO_2 and the warmer mode corresponds to the defrosted terrain.

Step 3: Dynamic Thresholding. We identify the characteristic "dip" (local minimum) between the two temperature modes, and select the corresponding temperature, T', as the threshold that distinguishes the polar cap from non-cap pixels. More specifically, we first identify the left and right peaks as local maxima in the histogram. We then identify the minimal point between them as the appropriate temperature threshold. In Figure 1(a), $T' = 175$ K. Finally, we filter out spurious detections by requiring that T' be in the range $[160, 210]$ K (based on domain knowledge).

Step 4: Cap Edge Identification. The CO_2 cap is not a

(a) Temperature histogram

(b) Temperature profile; north is to the left. Vertical lines mark the manual annotations of the defrosting zone as well as the cap edge found by our algorithm.
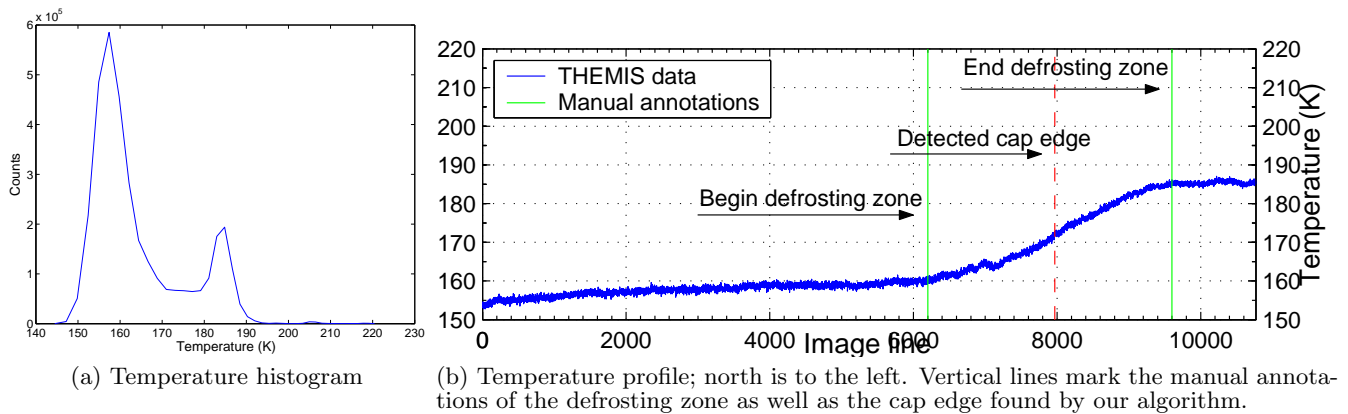
Figure 1: THEMIS data (band 9) for image I09779015.

discrete phenomenon with an abrupt "edge". Instead, it grows thinner with increasing distance from the pole, eventually becoming a thin layer of CO_2 frost, then isolated frost deposits, and then disappearing completely. Therefore, we define the cap "edge" as the point at which only 50% of the surface is covered in frozen CO_2. We apply the temperature threshold to the original image by marking each pixel that is colder than T' as belonging to the polar cap and each pixel that is warmer than T' as "non-cap". We then proceed from north to south and examine each line of the image, halting when we find a line that is less than 50% "cap". For image I09779015, we find the cap edge at latitude 61.05°N.

To provide further insight into this process, the temperature *profile* for image I09779015 is shown in Figure 1(b). This profile was generated by averaging all pixels in each image line to produce a single cross-track average temperature for that line. It shows the characteristic shape we observe in THEMIS images that contain the edge of the CO_2 cap: a sigmoid curve that transitions from a low temperature compatible with the presence of CO_2 ice to a temperature that is too warm to support CO_2 ice. The beginning of the defrosting zone, as annotated, occurs at about line 6200, at a temperature of 160 K, and it ends near line 9600, at 185 K.

4.2   Polar Cap Edge Detection: Results

To evaluate the detection rate and precision of our algorithm, we compared it to two independent methods for polar cap detection: a ground-based model that uses simultaneous, fully calibrated observations from an instrument with lower spatial resolution, and manual annotations of the THEMIS images.

4.2.1   Comparison to the TES Model

The Thermal Emission Spectrometer (TES), on-board Mars Global Surveyor, is also an IR camera in Mars orbit. TES observes at wavelengths ranging from 6 to 50 μm. Although TES has much lower spatial resolution than THEMIS, its temperature observations are much more reliably calibrated. The TES-based model is a 51-coefficient Fourier fit to cap edge locations identified in 60-km binned TES data, with a 1-sigma error estimate of 1.4 degrees of latitude [13]. For image I09779015, the TES model predicts that the cap edge to be at 61.96°N, which is 0.91 degrees north of our detection

and well within the margin of error for the TES model.

4.2.2   Manual Annotations

As a second source of independent validation, a student who was not involved in the algorithm development was trained to manually annotate the beginning and end of the defrosting zone in THEMIS images. The defrosting zone stretches from where the CO_2 ice begins to sublimate to the point at which no CO_2 remains. A total of 435 images were annotated in this fashion. Rather than looking at the temperature histogram or the images themselves, these annotations were generated after examining each image's temperature profile, as in Figure 1(b). Our manual annotator identified the beginning and end of this zone to the nearest 100 lines. Therefore, each annotation is specified ±10 km (1 line = 100 m). We interpret the midpoint as a first approximation to the edge of the cap. For image I09779015, this occurs at line 7900 (not explicitly shown in Figure 1(b)).

A natural question to ask is to what degree the manual annotations and the TES model predictions agree. We found high agreement in terms of deciding which images contained the polar cap (426 of 435), but less agreement about the exact location of the cap. The mean deviation between the manual annotations and the TES model was 2.07 degrees of latitude, or about 124 km, with a strong southward bias. That is, the manual annotations tended to indicate that the polar cap edge was further south than what the TES model would predict. Therefore, although we will evaluate against both standards (TES model and manual annotations), we rely more heavily on our comparison with the manual annotations, since they were derived from the same THEMIS data that our algorithm uses.

4.2.3   Performance

In terms of identifying which images contain the CO_2 polar cap edge, we find good agreement with both the TES model (96.3%) and with the manual THEMIS annotations (93.3%); see Table 1. The precision as measured against both standards is very high, as shown by the small number of false positives detected. Recall is somewhat lower in both cases, due to the larger number of false negatives.

For the 80 images in which the cap edge was detected, we

(a) Temperature histogram   (b) Temperature profile; north is to the left.
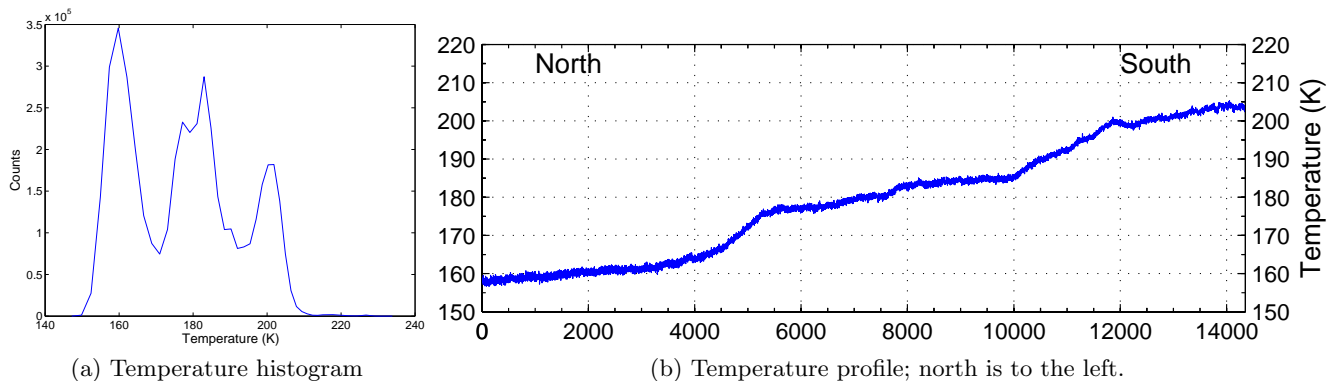
Figure 2: THEMIS data (band 9) for image I09626013.

Table 1: Agreement between our detections and two standards in identifying which THEMIS images contain the CO_2 polar cap edge (435 total images).

Standard	TES Model	Manual annotations
Recall	92.0%	86.4%
Precision	97.2%	94.3%
False pos.	4	8
False neg.	12	21
Agreement	419 (96.3%)	406 (93.3%)

also evaluated the degree to which the detected location of the cap edge matched both standards. The mean deviation between our detections and the TES model was 1.21 degrees (about 72 km, and within the TES model margin of error, 84 km). The mean deviation between our detections and the manual annotations was just 28 km; the estimated error in the manual annotations is 10 km. We also observe a bias in both cases; we tend to detect the cap edge slightly further south than where the TES model predicts it to be, and slightly north of where the manual annotations indicate that it is. This is consistent with our comparison to the TES model against the manual THEMIS annotations.

4.3 Scientific Discovery: Water Ice Annulus

During the course of our analysis of the polar images, we identified some anomalous images in which there are not two but three temperature modes present. In consultation with domain experts, we determined that the "middle" mode, with intermediate temperature values, is likely to correspond to a region that is covered by water ice as the CO_2 cap recedes north of it. Water ice exists at temperatures that are too warm to support CO_2 ice but too cold to be defrosted terrain. The existence of this water ice annulus had been posited based on modeling [7], and TES has seen some evidence for its existence [8], but this was the first time that supporting evidence at high spatial resolution was discovered. Identifying where water currently exists on Mars, and in what state, is a priority of the NASA Mars Exploration Program. As is often the case with large data sets, interesting facets of the data that would otherwise go unnoticed can be uncovered during the data mining process. We have written up this discovery for the benefit of the planetary science audience [15]; here, we relate the salient details as an example of the benefits of this kind of data mining.

A total of 197 polar images had either two or three temperature modes. Based on a manual examination of their temperature histograms, we found that 155 images had two modes and 42 images had three modes. Figure 2 shows the trimodal histogram and the corresponding temperature profile for one such image (I09626013). Comparing to Figure 1, we find that this image is qualitatively different in both histogram and profile; it has three modes, and instead of a single rise from cold to warm temperatures, there are two distinct "steps" in the profile (near lines 5000 and 11000).

We analyzed all 42 such trimodal images to identify the temperatures of each of their components using the k-means clustering algorithm [9]. For each image, we clustered all of the pixels with $k = 3$ and identified the mean temperature for each component. Although there is some overlap between the components, there is a clear separation in terms of the majority of observed temperatures. We interpret the three components, based on physical constraints, as follows:

	Temperature range	Mean temp.	Probable major constituent
Component 1	157–175 K	166 K	CO_2 ice/frost
Component 2	167–206 K	182 K	Water ice
Component 3	189–216 K	201 K	Defrosted terrain

In further analysis of the data, we found that the annulus tends to grow wider as spring advances and the seasonal cap recedes [15]. This finding had not been previously reported (or posited) and serves to increase our understanding of seasonal volatile cycling on Mars.

5. AEROSOL OPACITY ESTIMATION

The third algorithm was developed to identify high opacity atmospheric events. The opacity (or optical depth) is a measure of the amount of light removed by scattering or absorption as it passes through the atmosphere. Total opacity can be divided into components contributed by gases and various suspended particles. Here, we focus on two important components of the Martian atmosphere: dust and water ice particles, which form thin clouds. Atmospheric scientists are interested in the composition of the Martian atmosphere to better understand how gases, dust, and ice particles circulate on Mars. In addition, accurate estimations of the surface mineralogy from orbit depend on the ability to subtract out atmospheric constituents from the observations. Finally, on-board monitoring of the atmosphere can support

the early detection of dust storms and the identification of water ice clouds.

In previous work, Smith et al. analyzed fully calibrated THEMIS data from bands 3-8 [11]. The model was also informed by surface emissivity and an atmospheric temperature profile derived from simultaneous TES observations. They used an iterative least-squares method to derive opacity values for dust and water ice opacities. After analyzing a martian year's worth of THEMIS data and evaluating their model on synthetic spectra, they determined that the uncertainty associated with their aerosol estimates was about 0.04 or 10% of the total optical depth, whichever is larger.

The objective is to be able to identify high opacity events. Since the "background" atmospheric opacities for dust and ice vary with season, scientists should be able to specify an optical depth threshold that defines events of interest (for example, a dust τ that exceeds 0.8) based on the current season. Any images collected with an opacity above the specified limit constitute a detection. To address the issue of performance accuracy for detecting events, we first assessed how accurately we can estimate dust and water ice opacities of the Martian atmosphere using only uncalibrated THEMIS data.

5.1 Aerosol Opacity Estimation: Algorithm

Our goal was to build a regression model that maps THEMIS observations at different wavelengths to the dust and water ice optical depth values as computed by Smith et al. We focus on a framelet-based analysis here for several reasons. First, the training data is labeled on a per-framelet basis. In addition, aggregating pixels into framelets greatly reduces the computational cost of estimating opacity. Estimating opacity on a framelet basis provides a sufficiently find-grained result that satisfies the science goals of this mission. We also scaled the input data so that each band had a zero mean and unit standard deviation.

We used an SVM regression [6] approach to the problem. This model attempts to trade off a fit to the data with a "flatness" bias that provides better generalization properties (to new observations). Given a training data set composed of items x_i and associated opacities τ_i, the SVM regression problem is phrased as follows:

$$\max -\frac{1}{2}\sum_{i,j}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j)$$
$$-\epsilon \sum_i (\alpha_i^* + \alpha_i) + \sum_i \tau_i(\alpha_i^* - \alpha_i)$$

subject to

$$\sum_i (\alpha_i^* - \alpha_i) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C],$$

where α_i, α_i^* are Lagrange multipliers, $(x_i \cdot x_j)$ is the dot product of x_i and x_j, and ϵ is the tolerance associated with the regression fit. The output of the support vector machine, for a given observation x, is obtained by computing

$$f(x) = \sum_i (\alpha_i - \alpha_i^*)(x_i \cdot x) - b,$$

where b is a bias term that permits curves that do not pass

through the origin. If either α_i or α_i^* is greater than 0, then x_i is considered a support vector.

This formulation only penalizes the solution for errors that are greater than ϵ. For our experiments, we set ϵ to 0.01 and C to 50. It is possible to use the same method with a kernel function $\mathcal{K}$ that implicitly maps the input data into a higher feature space to permit nonlinear fits, so that the dot product is expressed in terms of the kernel function:

$$f(x) = \sum_i (\alpha_i - \alpha_i^*)\mathcal{K}(x_i, x) - b. \quad (1)$$

In our experiments, we used a Gaussian kernel ($\sigma = 0.1$) due to its superior results on this data set.

One way to reduce the cost of computing the output (opacity) for a new observation (framelet) is to construct a reduced-set SVM that approximates a given SVM with far fewer support vectors [2]. That is, instead of using s support vectors selected from the training set X as in Equation 1, we construct t ($t \ll s$) new vectors z_i, with coefficients β_i and bias term b', such that

$$f'(x) = \sum_i \beta_i \mathcal{K}(z_i, x) - b' \quad (2)$$

is as close to the output of the original SVM as possible. We use the reduced-set method proposed by Tang and Mazzoni [12], which yields more accurate approximations more efficiently than previous techniques. The reduced-set approach is what made this algorithm feasible for use on-board, as discussed in Section 6.

## 5.2 Aerosol Opacity Estimation: Results

We trained separate SVM regression models to estimate dust and water ice opacity. The total data set consists of 223,690 labeled framelets. We created a training set by arbitrarily selecting every 50th framelet (2209 framelets) and reserving the rest for testing (221,481 framelets). The original SVMs identified 1838 support vectors for the dust opacity estimator and 858 support vectors for the ice opacity estimator. We created reduced-set versions of each SVM that were limited to 40 support vectors. We evaluated each model in terms of the square root of the mean squared error (RMSE) as well as the mean error (Merr):

Method	Dust		Ice	
	RMSE	Merr	RMSE	Merr
SVM	0.086	-0.038	0.016	-0.004
Reduced-set SVM	0.087	-0.040	0.016	-0.004

First, we find significantly lower errors when estimating ice opacity than when estimating dust opacity. The large number of support vectors selected for dust estimation supports our intuition that this is a more difficult problem. Water ice tends to be easier to detect because atmospheric dust can be easily confused with surface dust, when observing from orbit around the planet. The RMSE for ice opacity estimation is well within the uncertainty associated with the labels (0.04), while the RMSE for dust opacity exceeds this value. However, it is still sufficiently accurate for detecting events of interest. The mean error numbers indicate that, on average, the SVM estimates tend to be lower than the true values. We
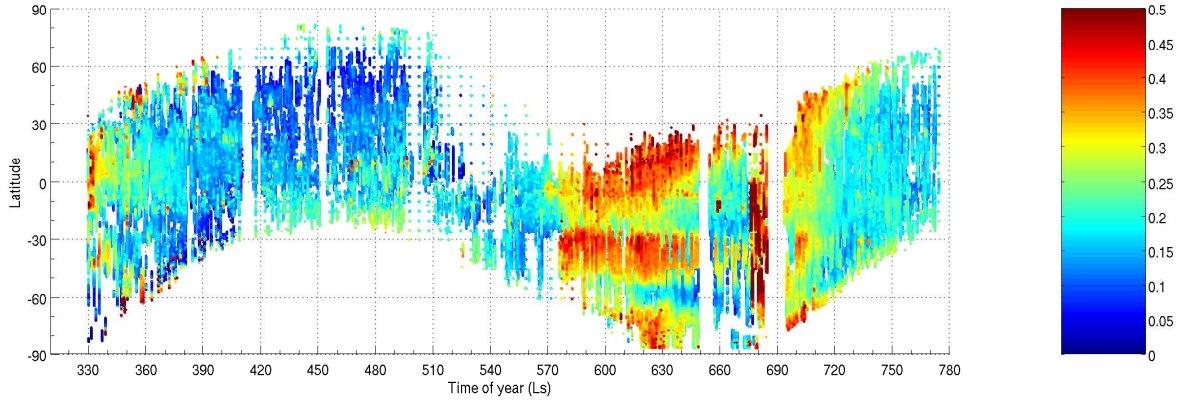
(a) Atmospheric dust optical depth as predicted by a Gaussian SVM regression.
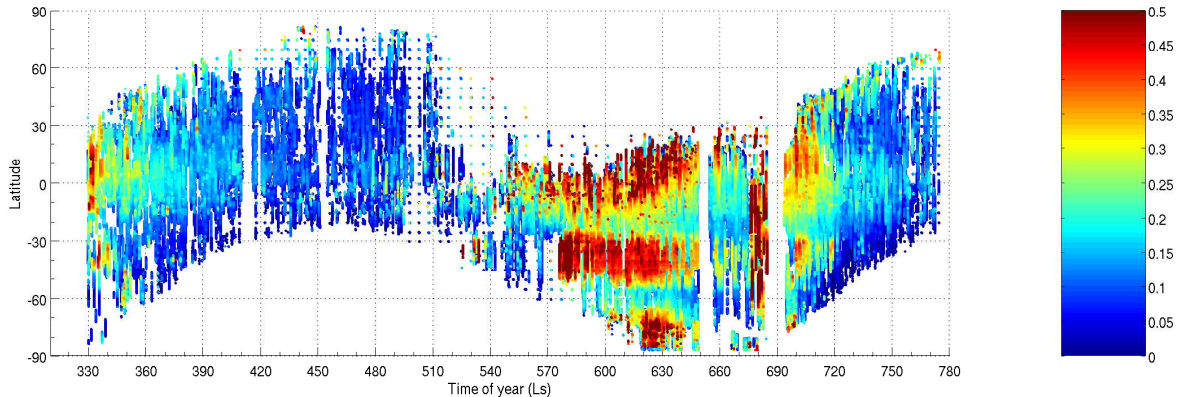
(b) "True" dust optical depth as reported by Smith et al.

Figure 3: Dust optical depths, predicted and "true" reference values. The x-axis indicates time of year (L_s). As in the paper by Smith et al. [11], opacity values are clipped to the ranges shown.

also find that the reduced-set SVM, while drastically reducing the memory consumption and processing time required to analyze a new framelet, does not significantly increase the error rate for either problem. We discuss the benefits of the reduced-set SVM further in the next section.

The τ predictions of the full SVM for dust opacity are shown in Figure 3(a), as a function of time of year and latitude. Time of year is commonly expressed as L_s, which refers to the planet's position in its orbit around the sun and varies from 0° to 360°. Here, we permit L_s to grow beyond 360 to show successive years on the same plot. These results match those of Smith et al. in Figure 3(b) quite closely, with the same dust event observed early on and three large storms appearing in around $L_s = 580°$. However, the magnitude of these events is slightly underestimated by the SVM, consistent with the mean error results above. The τ values predicted for the water ice opacity (Figure 4) provide an even better match to the reference values.

As described above, in an operational setting, scientists would be able to specify a minimum τ threshold that defines events of interest. A separate threshold could be specified for dust and water ice opacity analysis. A positive detection could be handled in a variety of ways, depending on the bandwidth available, ranging from a single bit indicating that a detec-

tion occurred up to the transmission of full observation that triggered the detection. We conducted experiments with a limited bandwidth scenario, in which only $x\%$ of the data that is collected can be transmitted. We specified a threshold of 0.4 for dust opacity and 0.2 for ice opacity. We evaluated the *hit rate* achieved for a given bandwidth limit as the ratio of transmitted framelets of interest to total framelets transmitted. If framelets are randomly selected for transmission, a constant baseline hit rate is achieved (see Figure 5). This hit rate is about 2% for dust events and 4% for water ice clouds. However, if we use the SVM regression method to estimate the opacity of each framelet, we can increase this hit rate dramatically. The benefits of this approach are most apparent when bandwidth is severely limited. This is the case for any event that is rare.

6. MOVING ALGORITHMS ON-BOARD

After the development of the algorithms and evaluation to validate that they meet accuracy performance requirements, they must be ported to a flight software environment and tested for computational resource usage. The on-board environment is constrained in a number of ways. All software intended to run on the spacecraft must run in a very limited memory footprint using only static or pre-allocated memory. The processor is also a carefully controlled resource. The
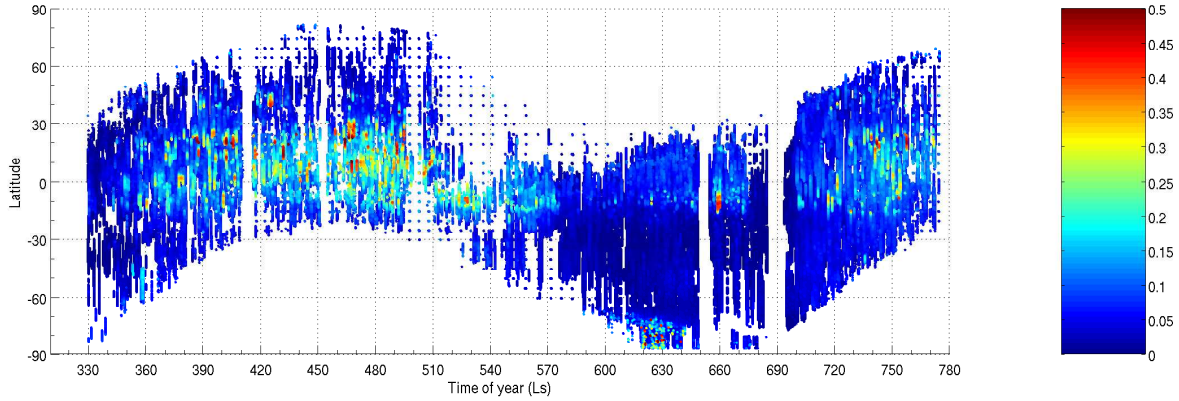
(a) Atmospheric water ice optical depth as predicted by a Gaussian SVM regression.
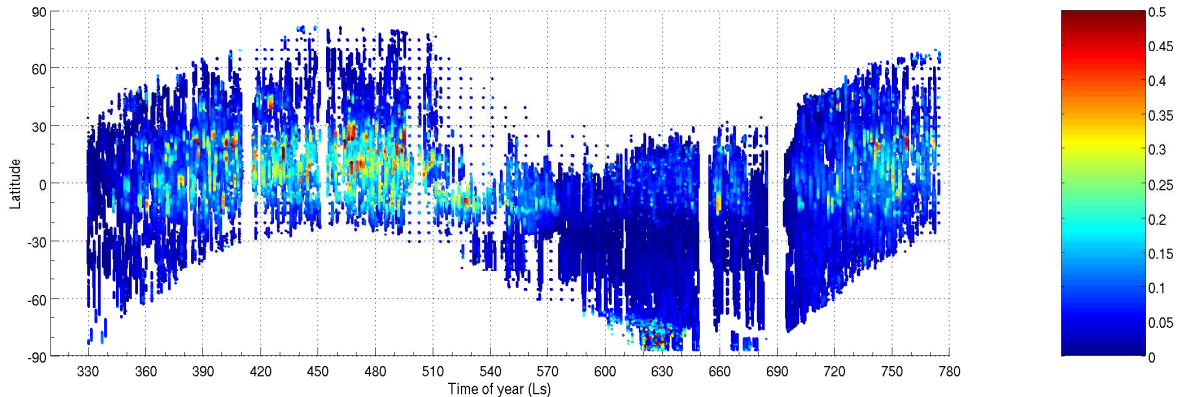
(b) "True" ice optical depth as reported by Smith et al.

Figure 4: **Water ice optical depths, predicted and "true" reference values. The x-axis indicates time of year** (L_s). **As in the paper by Smith et al. [11], opacity values are clipped to the ranges shown.**

algorithms must run efficiently and fast enough to keep up with the data acquisition rate of the instrument. During operations, machine learning algorithms running on-board will only be allocated a fraction of the total processing power. Runaway algorithms may be terminated or could result in the spacecraft entering a safe-mode which would disable it temporarily. It is therefore important to characterize the operation of the algorithms in an environment as similar to that on-board the spacecraft as possible.

6.1 Evaluation of Resource Requirements

The algorithms were initially developed and tested in the Matlab environment. Next, they were ported to C under Linux. The Linux versions of the algorithms were tested to ensure that they reproduced the results obtained under Matlab. Small changes were then required to port the algorithms to the VxWorks operating system running on the PPC750 testbed. Again, the algorithms were tested to ensure that they reproduced the same results in the new environment. The PPC750 testbed simulates a flight-like software configuration. It differs from a spacecraft in two ways: there is access to mass storage using file I/O, and we have control over 100% of the processor.

The algorithms were profiled for execution time and mem-

ory consumption on a Linux workstation (Workstation in Table 2) and on the PPC750 testbed (Testbed in Table 2). The Workstation machine was configured with a 1.793 GHz AMD Opteron Processor with 8 GB RAM, and the Testbed machine with a 150 MHz PPC750 with 128 MB of RAM. For the purposes of testing, time required for file I/O was not counted. We report execution speed for the thermal anomaly and polar cap detector algorithm in terms of the number of pixels per second that can be processed. We report the number of framelets per second for the aerosol opacity estimation algorithm. We tested both the full SVM and the reduced-set SVM on the Testbed, but only the full SVM on the Workstation. We considered running the reduced-set SVM on the Workstation unnecessary since the processing time is strictly linearly related to the number of support vectors. In Table 2, the performance of reduced-set SVM on the Workstation was estimated from the full SVM results.

The Mars Odyssey spacecraft has a RAD6000 processor running at 20 MHz. On-board analysis methods would only be allowed to use about 20% of the processor. They would have up to 40MB of heap memory available. The THEMIS instrument collects data at the rate of about 9600 pixels per second. The algorithms would be expected to run during and after data acquisition to determine if the collected data contains an event of interested should therefore be stored for
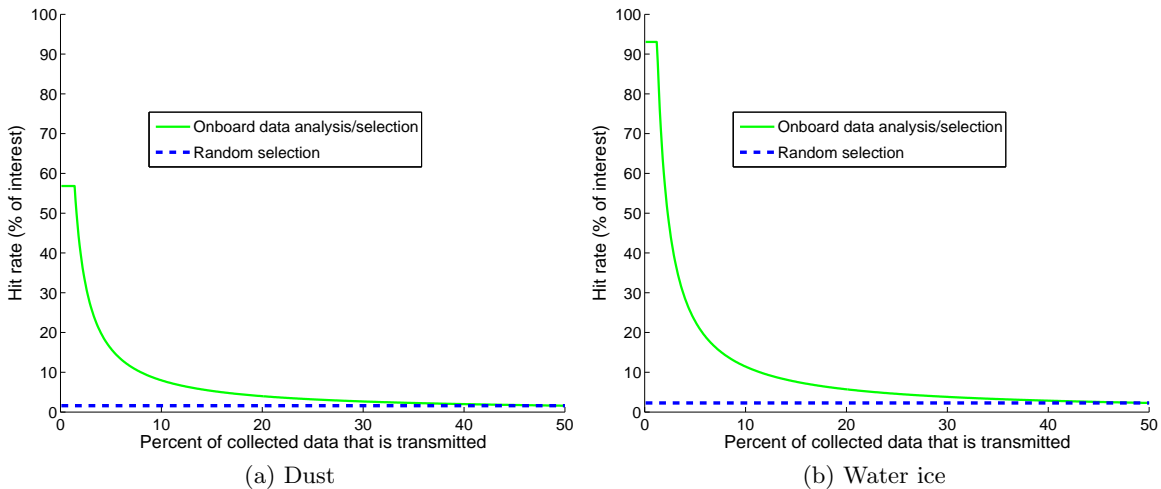
(a) Dust   (b) Water ice

Figure 5: The fraction of interesting framelets returned as a function of bandwidth constraints.

later transmission to Earth. Although the PPC750 Testbed is similar to the spacecraft, its processor is of a more recent generation and is approximately 10 times faster. Taking into account that on-board algorithms are only allotted only 20% of the processor, the Testbed is about 50 times faster. Therefore, we also show estimated processing speeds for each algorithm in a Spacecraft setting (Table 2). Our profiling results show that the algorithms can easily keep up with the output of the instrument even given the limitations of on-board processing power.

6.2 Lessons Learned

The successful implementation of machine learning in an operational system on-board a spacecraft requires addressing challenges that range from the analytical technical realm, to the fuzzy, philosophical domain of entrenched belief systems held by scientists and mission managers. Here we briefly discuss several practical lessons learned during this study.

First, the purpose of on-board science data analysis algorithms is to increase the mission science return. Therefore, to ensure that the results directly address mission needs, it is essential to work closely with domain scientists to understand the specific scientific problems they are addressing and how the use of on-board algorithms may help them achieve their goals. Working with the domain scientists must be at a tight level of interaction. They must provide information and define interestingness in ways that they may not be accustomed to as well as understand the practical limitations of an on-board algorithm. It often requires several iterations to arrive at well defined interest criteria that are practical for a detection algorithm. Generally, false alarms and missed detections have very different costs associated with them, and these factors must be addressed by the system.

The second important lesson is that the on-board data mining algorithms do not necessarily have to perform data analysis at the level of fidelity of a ground-based analysis algorithm. This is a key aspect of the problem that can be exploited so that the system can fit within the computational resource limits. For example, the science goal may be to characterize small dust storms. The on-board algorithm

need not be able to reliably identify all dust storm parameters that the scientist may want to know about the dust storm. The on-board algorithm need only identify the presence of a dust storm and indicate that this data should be marked as high priority to send to the ground for complete scientific analysis.

Third, as with most data mining systems that are to be deployed, it is important to start with simple methods and add complexity only when necessary. For example, the thermal anomaly detector uses a global threshold. As a result, it is not sensitive to small local thermal changes. It would have been more desirable to be able to have a local analysis algorithm. This was not selected due to three factors. First, no positive examples exist, because a true thermal anomaly has never been observed. Second, the actual occurrence of a thermal anomaly is very unlikely. The cost of developing a locally adaptive algorithm including testing and minimizing false alarms, without true positive examples, was too high relative to the likelihood of such an event actually occurring. These tradeoffs must be addressed for each science problem that an on-board analysis system seeks to solve.

7. CONCLUSIONS

There are a number of benefits to mining scientific data on-board a spacecraft including data prioritization, summarization, and reaction to dynamic events. This is an emerging paradigm and presents a significant change from the traditional ways of operating a spacecraft. Successful on-board data mining must meet the accuracy requirements provided by scientists while operating within the constrained on-board computing environment. We have presented three such algorithms for use on-board the Mars Odyssey spacecraft. These algorithms have all met the science requirements and processing requirements and are in the process of being integrated into the Odyssey flight software for future use on-board the spacecraft.

In addition to developing the algorithms themselves, we have also conducted a careful empirical study to assess the resource requirements and to determine whether they are realistic given the anticipated spacecraft computing environ-

Table 2: Resource requirements for all three data analysis methods; "pix" stands for "pixels" and "flts" stands for "framelets". The THEMIS instrument collects data at approximately 9.6 Kpix/sec (0.12 flts/sec).

Algorithm	Processing Speed			Memory in KB
	Workstation	Testbed	Spacecraft (est.)	(code segment)
Thermal anomaly detection	537.3 Mpix/sec	6.9 Mpix/sec	140 Kpix/sec	2
Polar cap edge detection	148.0 Mpix/sec	2.4 Mpix/sec	48 Kpix/sec	4
Dust opacity (full SVM)	5500 flts/sec	160 flts/sec	3.2 flts/sec	66
Ice opacity (full SVM)	12,000 flts/sec	350 flts/sec	7 flts/sec	30
Dust opacity (reduced SVM)	260,000 flts/sec (est.)	7560 flts/sec	151 flts/sec	2
Ice opacity (reduced SVM)	260,000 flts/sec (est.)	7570 flts/sec	151 flts/sec	2

ment. We have demonstrated that each algorithm falls well within the CPU and memory constraints. Finally, we have contributed a discussion of the lessons learned when working in this application area that can serve to guide future efforts to enhance the analysis capabilities of spacecraft.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] J. L. Bandfield, D. Rogers, M. D. Smith, and P. R. Christensen. Atmospheric correction and surface spectral unit mapping using Thermal Emission Imaging System data. *Journal of Geophysical Research*, 109(E10):E10008, 2004.

[2] C. J. C. Burges. Simplified support vector decision rules. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 71–77, 1996.

[3] R. Castano, D. Mazzoni, N. Tang, T. Doggett, S. Chien, R. Greeley, B. Cichy, and A. Davies. Onboard classifiers for science event detection on a remote sensing spacecraft. In *Proceedings of the Twelfth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 845–851, 2006.

[4] S. Chien, R. Sherwood, D. Tran, B. Cichy, G. Rabideau, R. Castaño, A. Davies, D. Mandel, S. Frye, B. Trout, S. Shulman, and D. Boyer. Using autonomy flight software to improve science return on earth observing one. *Journal of Aerospace Computing, Information, and Communication*, 2(4):196–216, April 2005.

[5] P. R. Christensen, J. L. Bandfield, J. F. Bell, N. Gorelick, V. E. Hamilton, A. Ivanov, B. M. Jakosky, H. H. Kieffer, M. D. Lane, M. C. Malin, T. McConnochie, A. S. McEwen, H. Y. McSween, G. L. Mehall, J. E. Moersch, K. H. Nealson, J. W. Rice, M. I. Richardson, S. W. Ruff, M. D. Smith, T. N. Titus, and M. B. Wyatt. Morphology and composition of the surface of Mars: Mars Odyssey THEMIS results. *Science*, 300:2056–2061, June 2003.

[6] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.

[7] H. Houben, R. M. Haberle, R. E. Young, and A. P. Zent. Modeling the Martian seasonal water cycle. *Journal of Geophysical Research*, 102(E4):9069–9083, 1997.

[8] H. H. Kieffer and T. N. Titus. TES mapping of Mars' north seasonal cap. *Icarus*, 154:162–180, 2001.

[9] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297, Berkeley, CA, 1967. University of California Press.

[10] D. E. Smith and M. T. Zuber. Seasonal changes in the icecaps of Mars from laser altimetry and gravity. In *Proceedings of the 13th International Workshop on Laser Ranging: Science Session*, 2002.

[11] M. D. Smith, J. L. Bandfield, P. R. Christensen, and M. I. Richardson. Thermal Emission Imaging System (THEMIS) infrared observations of atmospheric dust and water ice cloud optical depth. *Journal of Geophysical Research*, 108(E11):5115–5124, 2003.

[12] B. Tang and D. Mazzoni. Multiclass reduced-set support vector machines. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 921–928, 2006.

[13] T. N. Titus. Mars polar cap edges tracked over 3 full Mars years. In *Proceedings of the 36th Lunar and Planetary Science Conference*, Mar. 2005. Abstract #1993.

[14] K. L. Wagstaff, R. Castaño, S. Chien, A. B. Ivanov, E. Pounders, and T. N. Titus. An onboard data analysis method to track the seasonal polar caps on Mars. In *Proceedings of the Eighth International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2005.

[15] K. L. Wagstaff, A. B. Ivanov, T. N. Titus, and R. Castaño. Observations of the north polar water ice annulus on Mars using THEMIS and TES. *Planetary and Space Science*, 2007. In press.