The Astronomical Journal, 128:2965–2976, 2004 December

IDENTIFYING RED VARIABLES IN THE NORTHERN SKY VARIABILITY SURVEY[1]

P. R. Woźniak, S. J. Williams, W. T. Vestrand, and V. Gupta
Los Alamos National Laboratory, Mail Stop D436, Los Alamos, NM 87545;
wozniak@lanl.gov, steven@lanl.gov, vestrand@lanl.gov, varsha@lanl.gov
Received 2004 June 2; accepted 2004 August 26

ABSTRACT

We present a catalog of 8678 slowly varying stars with near-infrared colors corresponding to the evolved asymptotic giant branch population. Objects were selected from the Northern Sky Variability Survey (NSVS) covering the entire sky above declination $\delta = -38^\circ$ in a single unfiltered photometric band corresponding to a V-band magnitude range of $\sim$8–15.5 mag. After quality cuts, the number of measurements for a typical star is approximately 150, but it ranges up to $\sim$1000 for high-declination stars. We show that the use of support vector machines, a modern machine-learning algorithm, can reliably distinguish Mira variables from other types of red variables, namely, semiregular and irregular. We also identify a region of parameter space that is dominated by carbon stars. Our classification is based on period, amplitude, and three independent colors possible with photometry from the NSVS and the Two Micron All Sky Survey. The overall classification accuracy is $\sim$90% despite the relatively short survey baseline of 1 yr and limited set of features. There are 6474 stars in our sample without identifications in the General Catalogue of Variable Stars, which, as such, are most likely new discoveries. Period-amplitude and period-color diagrams of both our previously known and newly identified Mira stars are in good agreement with published studies based on smaller samples.

Key words: catalogs — stars: AGB and post-AGB — stars: variables: other

Online material: machine-readable table

1. INTRODUCTION

Variable stars in the red and luminous part of the H-R diagram are often collectively referred to as red variables. They are mostly asymptotic giant branch (AGB) stars of K and M spectral types and are traditionally classified into Mira (M), semiregular (SR), and slow irregular (L) variables, although some recent work suggests that there are also variables at the top of the first giant branch (e.g., Ita et al. 2002). Pulsation instability is the most likely explanation of variability in those objects (e.g., Keeley 1970). Much of their importance comes from the fact that Mira variables and Mira-like semiregulars obey period-luminosity relations and therefore can be used as distance indicators (Feast 2004). There are also indications that some apparently irregular variables are actually multiperiodic and also form period-luminosity sequences (Wood et al. 1999; Mattei et al. 1997; Wray et al. 2003). AGB stars are intrinsically bright, and therefore even a relatively shallow survey to the flux limit of $\sim$16 mag can detect them throughout the Galaxy. Red variables are therefore good tracers of intermediate-age to old populations and have been used to study kinematics of the Galaxy (e.g., Kharchenko et al. 2002; Feast & Whitelock 2000; Luri et al. 1996). Since they are undergoing significant mass loss, they play an important role in the enrichment of the interstellar medium (Willson 2000).

The primary means to find and initially classify red variables are time-resolved photometry and light-curve morphology. Temporal data also constrain stellar pulsation models (see Reid & Goldston 2002) and give insights into shock wave propagation in the rarefied atmosphere of an AGB star (Maffei & Tosti 1995)—hence the need for homogeneous sets of precise photometry using modern CCD techniques. The fourth edition of the General Catalogue of Variable Stars (GCVS; Kholopov et al. 1998) lists 19,043 objects in all categories of red variables, $\sim$15,000 reaching magnitude $V = 15.5$ or brighter at maximum. However, at magnitude $V = 13.0$ the Mira subsample is less than 10% complete (Kharchenko et al. 2002). Good light curves for large samples of red variables have been hard to find. Visual observations by members of AAVSO and other amateur observers around the world are often the only source of monitoring information, especially over long time baselines.

The situation is rapidly improving with the advent of massive photometric monitoring experiments. Daily sampled light curves covering up to a decade for a few hundred thousand variable stars of many types in the Galactic bulge region, LMC, and SMC have been published by microlensing surveys (OGLE, MACHO, MOA, EROS; e.g., Paczyński 2000 and references therein). Major recent additions to the data pool on the largest spatial scales are the Northern Sky Variability Survey (NSVS;[2] Woźniak et al. 2004) and the All Sky Automated Survey (ASAS; Pojmański 2002). Together they cover the whole sky in the intermediate magnitude range 6–15, which is easily accessible to spectroscopic and astrometric follow-up.

With this paper we release a catalog of 8678 red variables from NSVS comprising broadband optical light curves from NSVS and near-infrared colors from the Two Micron All Sky Survey (2MASS).[3] After discussing the data set and the selection

[1] Based on observations obtained with the ROTSE-I Robotic Telescope operated at Los Alamos National Laboratory.

[2] This publication makes use of the data from the Northern Sky Variability Survey created jointly by the Los Alamos National Laboratory and the University of Michigan. The NSVS was funded by the Department of Energy, the National Aeronautics and Space Administration (NASA), and the National Science Foundation (NSF).

[3] This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the NASA and the NSF.

of variables, we explore the possibility of applying machine learning to the classification of variable stars and selecting Mira-like stars in particular. Proliferation of large, multidimensional astronomical data sets is stimulating work on new methods and tools to handle the challenge (e.g., Banday et al. 2001). In many cases machine learning can provide the solution.

2. DATA

A detailed description of NSVS was published by Woźniak et al. (2004). The survey was conducted in the course of the first-generation Robotic Optical Transient Search Experiment (ROTSE-I; Akerlof et al. 2000). The SkyDOT database provides a convenient browsing and search interface for NSVS and was used extensively in the preparation of our catalog.[4]

The main issues to keep in mind when interpreting NSVS data are related to the very wide field of view of the system and the broad, unfiltered spectral response. The spectral window extends from the mid-B to mid-I bands, and the effective wavelength is that of the R band. The ROTSE-I system had four comounted f/1.8 Cannon cameras, each with the $8^\circ \times 8^\circ$ field of view covered by a $2\mathrm{k} \times 2\mathrm{k}$ CCD detector. The resulting pixel size of 14.″4 significantly limits the spatial resolution of NSVS. With the field of view this large, true sky brightness and airmass gradients, as well as nonuniform thin clouds, limit the accuracy of photometric calibration. While the final photometry is tied to the Johnson V magnitude scale of the Tycho catalog, using Tycho $B - V$ colors in the process, there is no instrumental color information. Therefore, some irreducible systematic effects remain when the data collected in adjacent fields are compared, mostly affecting objects of unusual colors. Nevertheless, the internal consistency of light curves is very good, with rms photometric scatter for bright constant stars in a median field at the level of 0.02 mag. NSVS covers a flux range roughly corresponding to V-band magnitudes between 10 and 15.5, with some measurements as bright as 8 mag because of vignetting and shorter bright-time exposures.

To facilitate classification of catalog variables we make use of near-infrared $J - H$ and $H - K_s$ colors from 2MASS (All-Sky Point Source Catalog, Release 2003 March 25). Positional cross-correlations between 2MASS and our data are discussed in § 3.3. Objects in common between our sample and the sample of red variables with known types from the fourth edition of the GCVS (Kholopov et al. 1998) are used in § 4 as a training set for classification.

3. OBJECT SELECTION AND FEATURES

3.1. *Searching for a Slow Variability Pattern*

The selection process starts with almost 2×10^7 individual light curves in the NSVS data set. To reduce the number of objects to be considered as early in the process as possible, the first few cuts were performed on aggregate light-curve parameters. Only objects with a minimum of 100 observations total were considered. We required $\sigma > 0.1$ mag and $\sigma/E > 5.0$, where σ is the rms scatter and E is the median of error estimate of all "good" magnitude measurements without problem flags (see Woźniak et al. 2004 for details). This procedure selects 98,908 candidate light curves.

The next step requires extracting light curves for all candidates and aims at rejecting rapid variations through a procedure similar to the analysis of variance (AOV; Schwarzenberg-Czerny 1989). After binning the light curve into 15 day bins, the mean m_i and the standard deviation s_i of all points in each ith bin are calculated. Only bins with more than five points and objects with more than five valid bins are considered for further analysis. The AOV ratio $R_{\mathrm{AOV}} = (\langle m^2 \rangle - \langle m \rangle^2)^{1/2} / \langle s \rangle$ is the ratio of the standard deviation of all means to the mean of all standard deviations. As shown in Figure 1 this ratio separates stars with light curves correlated on timescales of 2 weeks or more from short-period variables, other fast variables, and occasional artifacts. There were 11,473 light curves with an AOV ratio above 1.6, our adopted threshold.

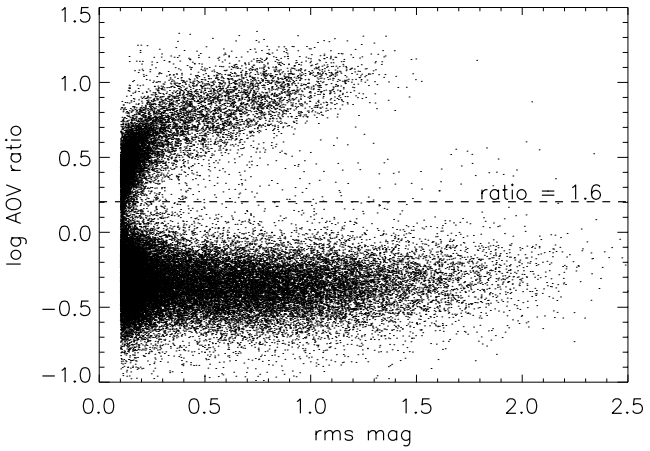

FIG. 1.—Initial separation of slow variables employing the AOV ratio, which is plotted vs. rms magnitude scatter.

Only 9371 of these variables turned out to be distinct objects, as some fraction of objects are detected in overlapping fields, especially at high declinations where tiling the sphere with fixed size frame becomes redundant and, at the same time, temporal coverage is very good. We assumed that objects closer than 14.″4 (ROTSE-I pixel size) are two detections of the same object. Before merging the data for multiple references to the same objects, some light curves required up to a few percent adjustment of the zero point because of the effects mentioned in § 2. From this point we continue with 9371 distinct objects with merged light curves, maintaining the reference to the origin of individual points in the final catalog.

3.2. *Periods and Amplitudes*

The 1 yr time span of NSVS makes it difficult in numerous cases to estimate amplitudes and periods precisely. Relations involving $\log P$ are not strongly affected as long as periods are known to 20% or better. The period distribution of known Mira variables peaks around 1 yr and extends to periods of about 3 yr. SR variables tend to have shorter timescales, but strictly speaking none of these types is truly periodic. Therefore, meaningful values for periods could be obtained by fitting a single sine wave with free amplitude and phase at trial periods ranging from 10 to 730 days with 1 day resolution and selecting the value resulting in the best fit. For a small percentage of stars this value is the upper boundary at 730 days as a result of insufficient data, in which case we can safely assume that the period is unknown but long. Examples of light curves in our sample and corresponding fits are given in Figure 2.

The total amplitude from our single sine-wave fits becomes unstable in cases when one of the extrema was not observed, even when the period is well determined from two extrema of the same type. We get an acceptable estimate for the amplitude by simply taking the total observed magnitude range as the

[4] Available at http://skydot.lanl.gov.

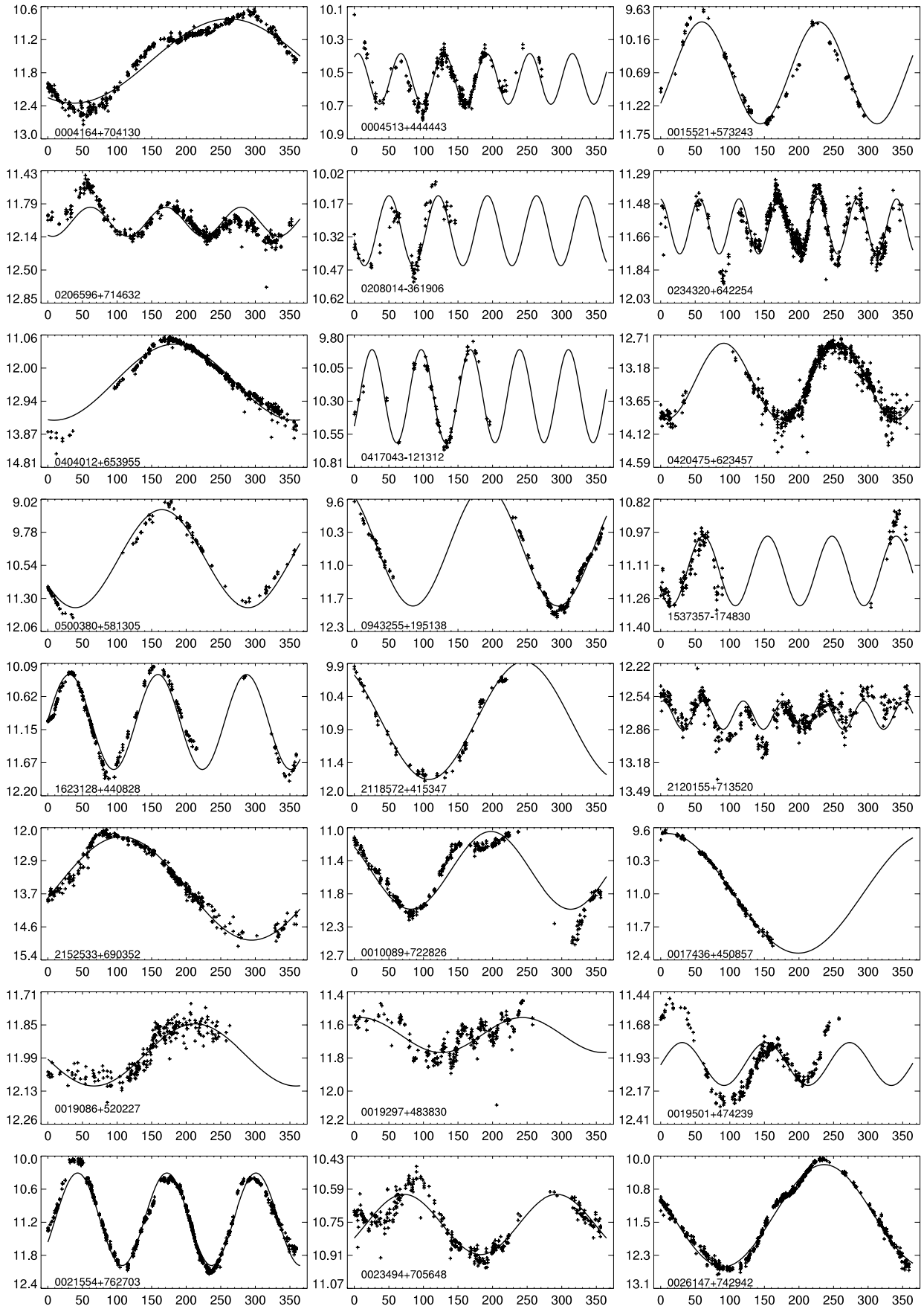

FIG. 2.—Example catalog light curves and corresponding fits (§ 3.2) for objects representative of the range of amplitudes, periods, data quality, and coverage.

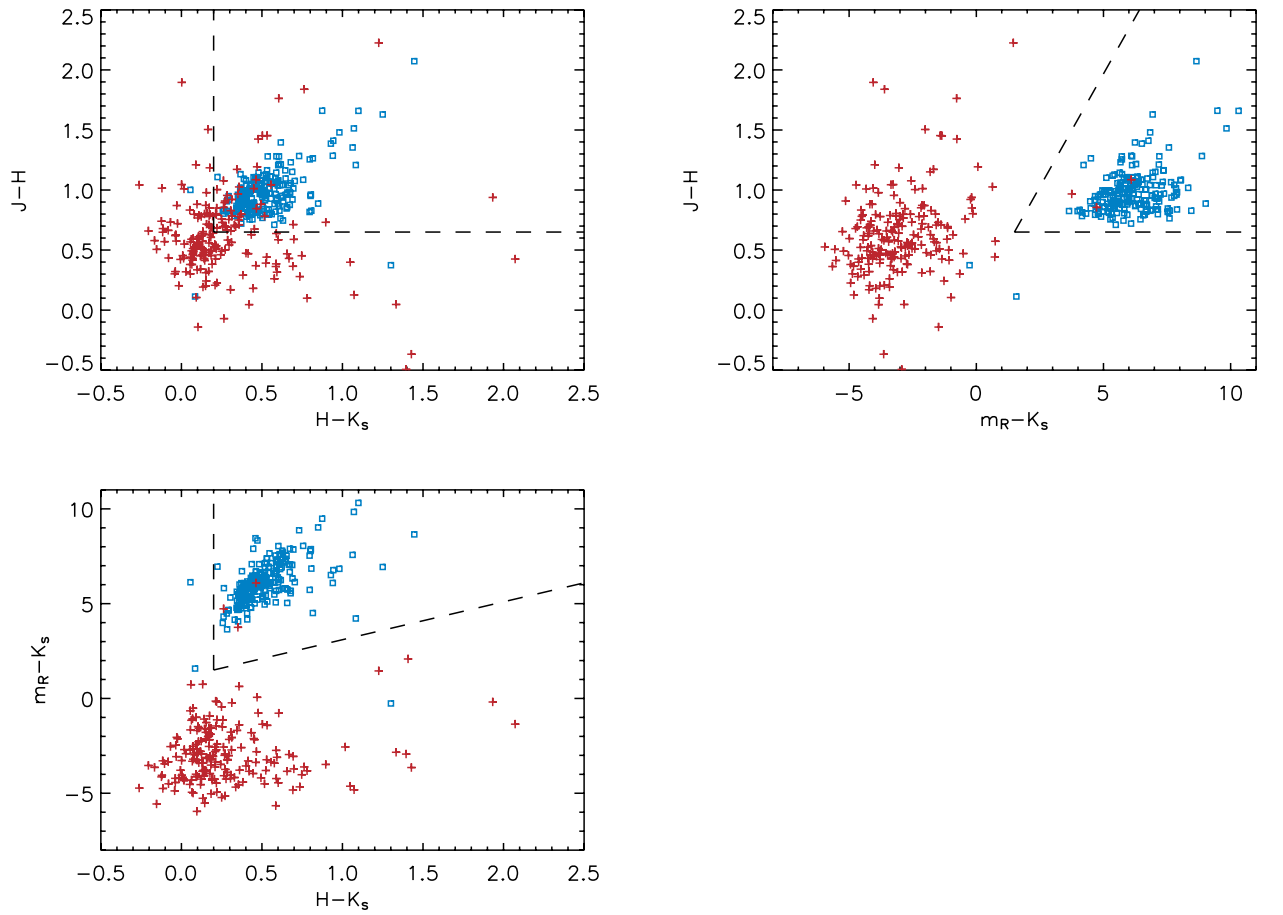

FIG. 3.—High-probability matches (*blue squares*) and purely spurious matches (*red plus signs*) to the 2MASS catalog in color space. Each group is represented by 200 objects. The dashed line denotes the boundary of the region adopted for identification of red variables.

difference between the third brightest and third faintest recorded magnitudes. For poorly observed objects this is usually an underestimate of normal behavior and will cause some Mira variables to be misclassified as SR+L. This is acceptable when a reasonably clean sample of Mira variables is required but may cause problems in other applications.

The amplitudes of AGB variables decrease toward the red part of the visual-infrared spectrum (Cioni et al. 2001; Whitelock et al. 2000; Eggen 1975), and for the ROTSE-I band m_R they are already reduced by a factor of a few compared with the V-band amplitudes. This fact is frequently used, especially in the infrared, to draw conclusions from colors formed out of nonsimultaneous measurements or taken at random phases. The typical variability amplitude of AGB stars in the K band is roughly $1/10$ of that in the visual.

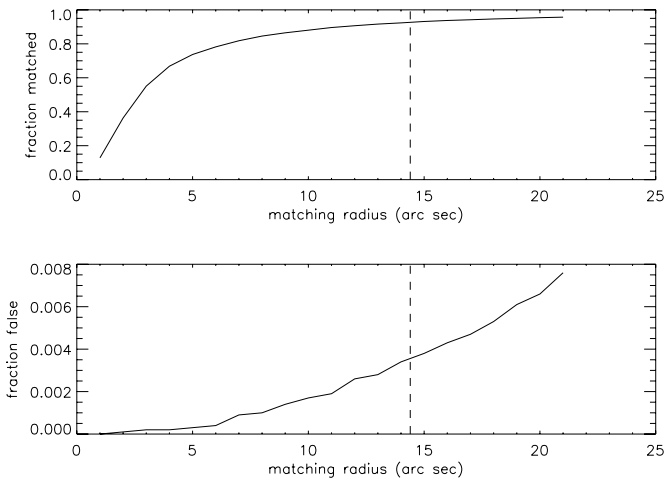

FIG. 4.—Rate of successful (*top*) and spurious (*bottom*) matches to 2MASS objects as a function of the tolerance radius.

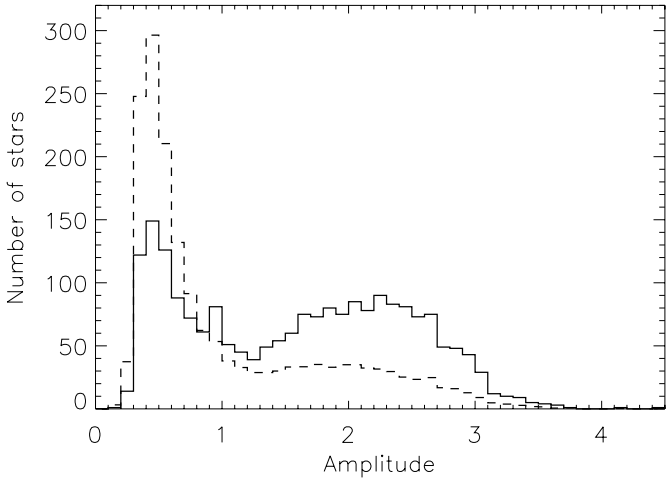

FIG. 5.—Amplitude distribution of stars in the training set (*solid histogram*) and all catalog stars (*dashed histogram*). Bimodality is weaker in the larger data set because of a much larger proportion of irregular variables. The dashed line was scaled down by a factor of 5.

TABLE 1
CONFUSION MATRIX FOR THE TWO-CLASS PROBLEM

	PREDICTED CLASS	
ACTUAL CLASS	M (%)	SR+L (%)
M	91.3	8.7
SR+L	16.8	83.2

3.3. *Colors*

The use of the near-infrared colors greatly improves the fidelity of red variable identification. We cross-correlate positions of preselected NSVS objects with the 2MASS survey, which is an excellent source of simultaneous JHK_s photometry. Using combined data we can form three independent colors: $m_R - K_s$, $J - H$, and $H - K_s$, where m_R is the median magnitude of the NSVS light curve.

Because of the large difference between positional accuracy and depth of NSVS versus 2MASS, maximizing the number of successful matches requires the use of additional information. A tolerance radius of about 14.″4, which is appropriate for NSVS, results in numerous random matches in 2MASS. However, the fact that NSVS stars are bright and the expected colors of our variables are very red allows us to easily spot most random coincidences. In Figure 3 we show the colors for variables with very secure matches (3″ radius) and a set of random matches. The random matches are generated from a spatial distribution following that of our actual data set. It is not surprising that the broadband $m_R - K_s$ color has the most discriminating power. Dashed lines indicate the boundary of the final adopted locus for "true" matches. Please note that the rejection of the blue tip of the region populated by true matches is intentional. Objects found in this region are primarily yellow giants of the SRD type with spectral class typically earlier than K. They are not very numerous, and at the same time there are many random matches with similar $J - H$ and $H - K_s$ colors.

In the case of multiple candidates within the extended 14.″4 identification radius, we adopt the closest 2MASS object. This procedure yields 8678 identifications. Only 51 of those had two candidates of the right color, and two had three such candidates. To better quantify our identifications, we repeated the process for a range of tolerance radii up to 21″ (∼1.5 ROTSE-I pixels). Figure 4 shows the number of positive matches (*top*) and the corresponding number of matches using random positions (*bottom*). Our final adopted matching radius (*vertical line*) admits fewer than 0.4% of purely random spatial coincidences that happen to match a 2MASS star of the desired color.

4. CLASSIFICATION

The analysis discussed in previous sections leaves us with five features to use in the classification of our red variables: period (P), amplitude (A), and $m_R - K_s$, $J - H$, and $H - K_s$ colors. Experiments with other features and different definitions of the above features did not improve the results. We decided to use support vector machines (SVMs) to handle

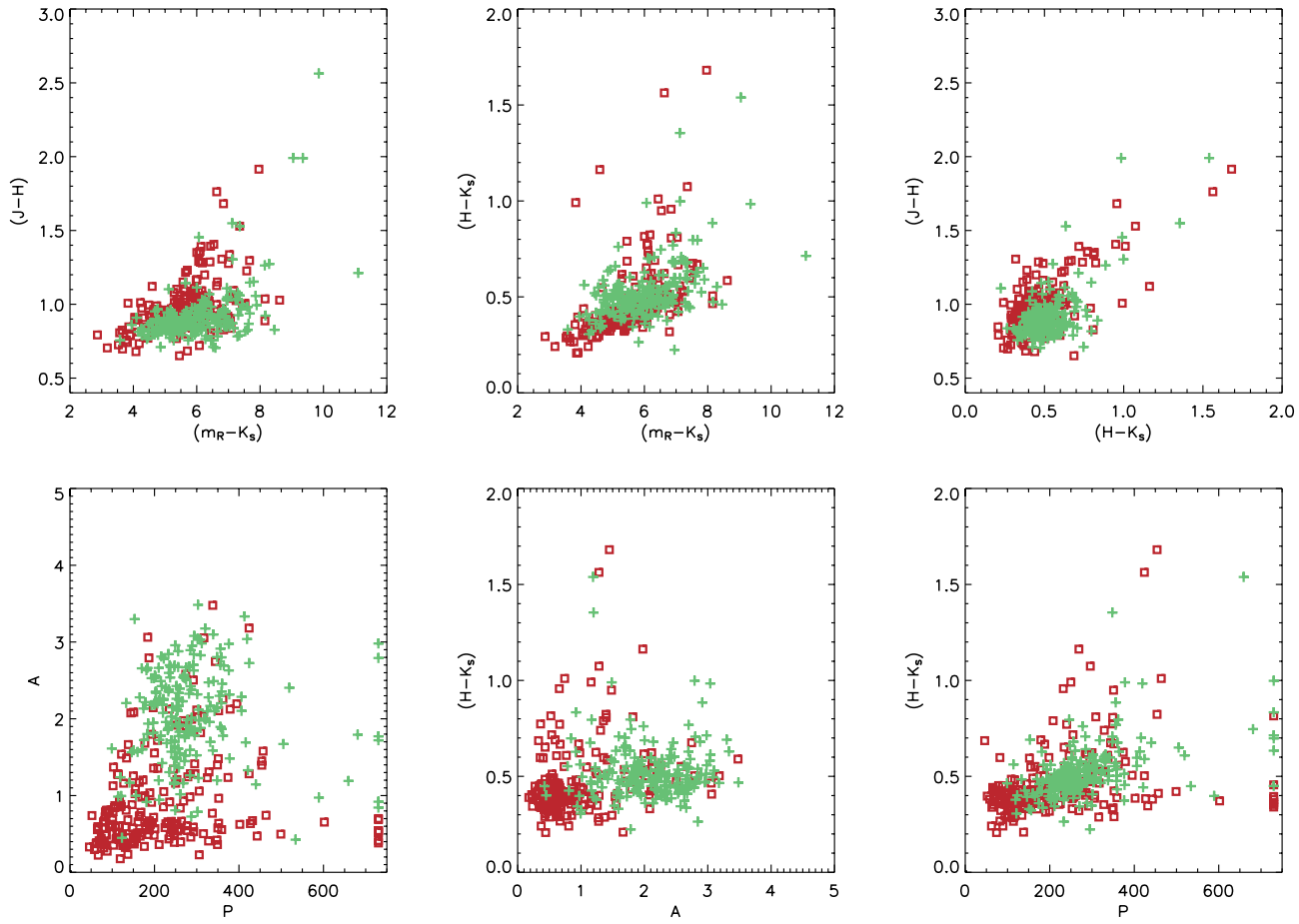

FIG. 6.—Training set in various projections of the input feature space for a two-class problem of separating Miras (M) from other long-period variables (LPVs; SR+L). Classes M and SR+L are shown with plus signs and squares, respectively. For clarity, only up to 200 objects per class are shown.

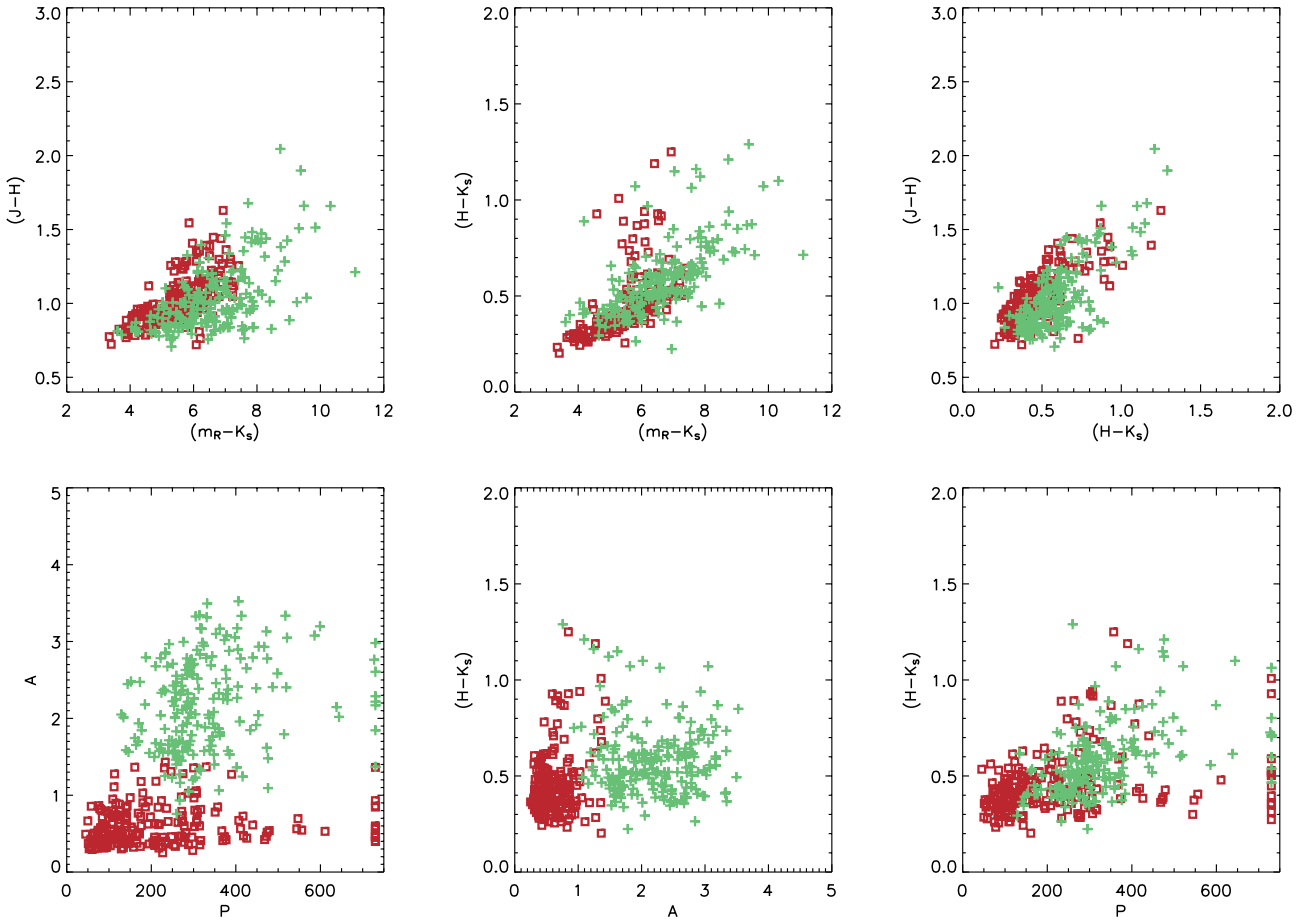

FIG. 7.—All classified stars in various projections of the input feature space for a two-class problem. Symbols are the same as in Fig. 6. For clarity, only up to 200 objects per class are shown.

smooth, overlapping distributions of unknown shape in a five-dimensional space. This algorithm is more objective and efficient than the traditional visual inspection. It eliminates the setting of arbitrary class boundaries in low-dimensional (typically two-dimensional) projections. A thorough introduction of SVMs is beyond the scope of this paper. We refer the interested reader to texts by Vapnik (1998) and Cristianini & Shawe-Taylor (2000). Examples of astronomical uses of SVMs include Woźniak et al. (2001), Humphreys et al. (2001), and Zhang & Zhao (2003). Below we only summarize several important points of relevance to our study.

4.1. *Support Vector Machines*

SVMs are a state-of-the-art method of supervised learning that requires a training set of data with known class membership. The result of the training run is a concrete classifier that can be applied to new data with unknown class membership. With the clever use of the kernel functions to transform the data into a high-dimensional feature space, SVMs are capable of finding highly nonlinear class boundaries using only hyperplanes. SVMs are inherently resistant to overfitting. The objective of the SVM is to maximize the so-called margin, a generalized orthogonal distance between the class boundary and points closest to the boundary on both sides. For a given set of input parameters, SVMs guarantee that the final result represents the global minimum of the objective function rather than one of the local minima. This is one of the reasons for the excellent generalization properties of SVMs on previously unseen data.

We used the publicly available LIBSVM implementation of SVMs (Chang & Lin 2001), which, for a N-class problem, performs a voting procedure using results from $N(N-1)/2$ binary classifications. The package offers so-called soft-margin SVMs capable of working with data that are not fully separable into a hyperplane in the n-dimensional space of transformed features. This capability is very important in the presence of noise, where the simpler maximal-margin machine usually breaks down. For a Gaussian kernel function the training algorithm employs two parameters, the width of the Gaussian kernel and the amplitude of the penalty term for misclassification, known respectively as $\gamma = 1/2\sigma^2$ and C. They are somewhat correlated and measure the level of coupling between the data vectors sensed by the algorithm (γ) and how hard SVMs should try to avoid misclassification by making the boundary more flexible (C).

4.2. *Finding Mira Variables*

For the purpose of training the machine and obtaining the classifier, we use a subset of our red variables in common with the GCVS and previously classified as M, SR, or L type. There are 2095 such stars, all uniquely identified within 14.″4 (1 ROTSE-I pixel). Because the distance in a multidimensional feature space is used to derive the classifier, rescaling of the input features usually improves results. We replaced the periods

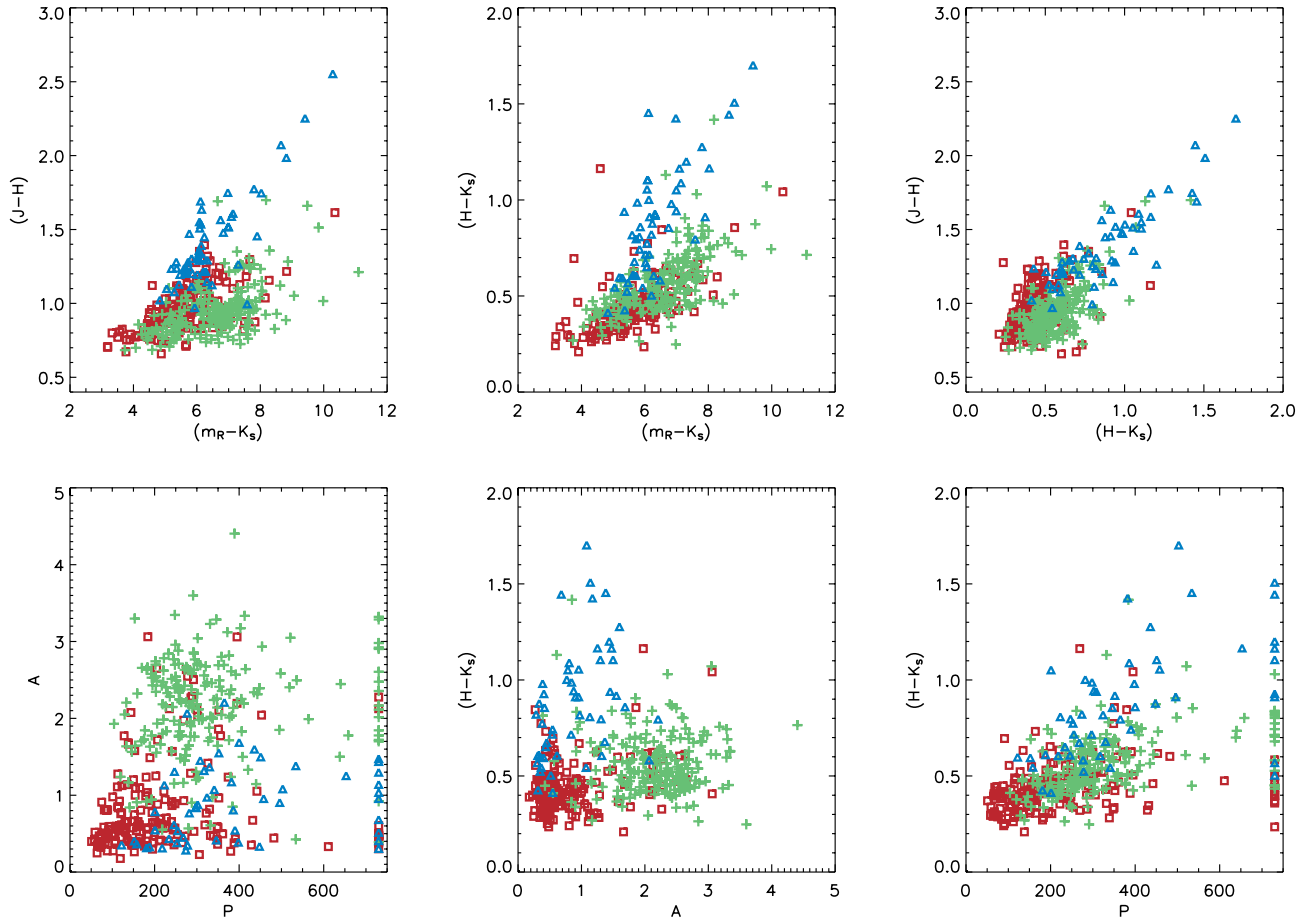

FIG. 8.—Training set in various projections of the input feature space for a three-class problem of separating Miras (M) from other LPVs (SR+L) and carbon stars (C). Classes M, SR+L, and C are shown with plus signs, squares, and triangles, respectively. For clarity, only up to 200 objects per class are shown.

with their logarithm and renormalized the distribution of each feature vector to zero mean and unit variance.

The first classification problem we attempted to solve was a three-class problem with types M, SR, and L. All classifications in this category suffered from significant confusion between SR and L types at the level of 30%. The classification was sensitive primarily to the gap in the bimodal amplitude distribution of the training set (Fig. 5), as previously discussed by Mattei et al. (1997), who considered AAVSO data averaged over many cycles only for M and SR stars from GCVS (see also Mennessier et al. 1997 and Whitelock et al. 2000). This gap gives the best handle on separating Miras from semiregulars at $P > 80$ days. As a result of long-term variations in red variables, this gap is not visible in GCVS data because it lists global historic maxima and minima. There is a slight progression of colors toward the red as one moves from L through SR to the M stars of the training set; however, the overlap between classes remains very large.

Unfortunately, statistical dereddening of colors does not change this situation. With the information at hand, we could only deredden the colors by assuming that our variables are standard candles of absolute magnitude $M_R = 0.5$ and employing an iterative procedure analogous to that of Feast et al. (1990). However, we used the total extinction from the map of Schlegel et al. (1998). While the distribution of dereddened colors is noticeably tighter and slightly shifted to the blue compared with the apparent colors, it has a negligible effect on the final classification. The results are also insensitive to the assumption of constant M_R and its precise value. Our photometric data therefore do not permit a reliable separation of variables of types SR and L.

Forced to settle on a two-class problem, we next classified Mira variables (M) versus "other" variables (SR+L). After some investigation of the success rate with various input parameters, we selected $\gamma = 0.04$ and $C = 1.0$. The confusion matrix $\mathcal{M}_{ij}$ for our sample, which measures the fraction of training instances in class i assigned to class j, is given in Table 1. Figures 6 and 7 present training data and the full set of classified data for various projections of the input space. The final classification accuracy for this problem was about 87%. Both the confusion matrix and the accuracy were estimated using a tenfold cross-validation scheme, where the SVM is repeatedly trained using 90% of the data, and its accuracy is estimated using remaining 10% of the sample. This estimate is a better predictor of performance on new data than classification

TABLE 2
CONFUSION MATRIX FOR THE THREE-CLASS PROBLEM

	PREDICTED CLASS		
ACTUAL CLASS	M (%)	SR+L (%)	C (%)
M	95.1	3.5	1.4
SR+L	14.6	82.4	3.0
C	4.5	6.5	89.0

TABLE 3
SUMMARY OF SVM CLASSIFICATION

Problem	Accuracy (%)	Parameters		Number of Objects					
				Training Data			All Data		
		γ	C	M	SR+L	C	M	SR+L	C
Two-class	87	0.04	1.0	1221	874	...	2565	6113	...
Three-class	90	0.10	1.0	417	177	53	2276	5719	683

accuracy on the training set. The best results were obtained by resampling the training set to achieve equal weighting of all classes. This was particularly important in the next attempt to identify carbon stars, which are rare in the GCVS and our training set.

4.3. *Carbon Stars*

In various color-color projections of the training data one notices a fairly distinct tail of points extending toward the reddest part of those diagrams. This tail corresponds to the $J - H$, $H - K$ locus of carbon stars discussed in Whitelock et al. (2000) and also Bessell & Brett (1988). There are only 647 stars with spectral information among M, SR, and L types from the GCVS that were also selected in NSVS. Of these, 53 have carbon spectra (regardless of variability type). Indeed, Figure 8 confirms that the subsample known to have carbon spectra almost exclusively falls within the red tail, while objects with other spectral types rarely overlap with the tail.

Consequently, we trained a classifier for a three-class problem with types M, C, and SR+L for Mira, "carbon," and "other." Only stars with known spectra were used in this training run. We also applied reweighting of the samples to reduce systematics due to the dissimilar number of available objects between the types. Despite the larger number of classes, allowing for a third class increased the final accuracy to about 90% (using $\gamma = 0.1$ and $C = 1.0$). For both two-class and three-class problems the variance of all cross-validation runs was about 1%. We interpret the third detected class as a group of carbon stars. Our derived confusion matrix for this classification is shown in Table 2. Table 3 lists the number of objects available for training in various classes, the input parameters, and the resulting classifications for both two-class and three-class runs. Projections

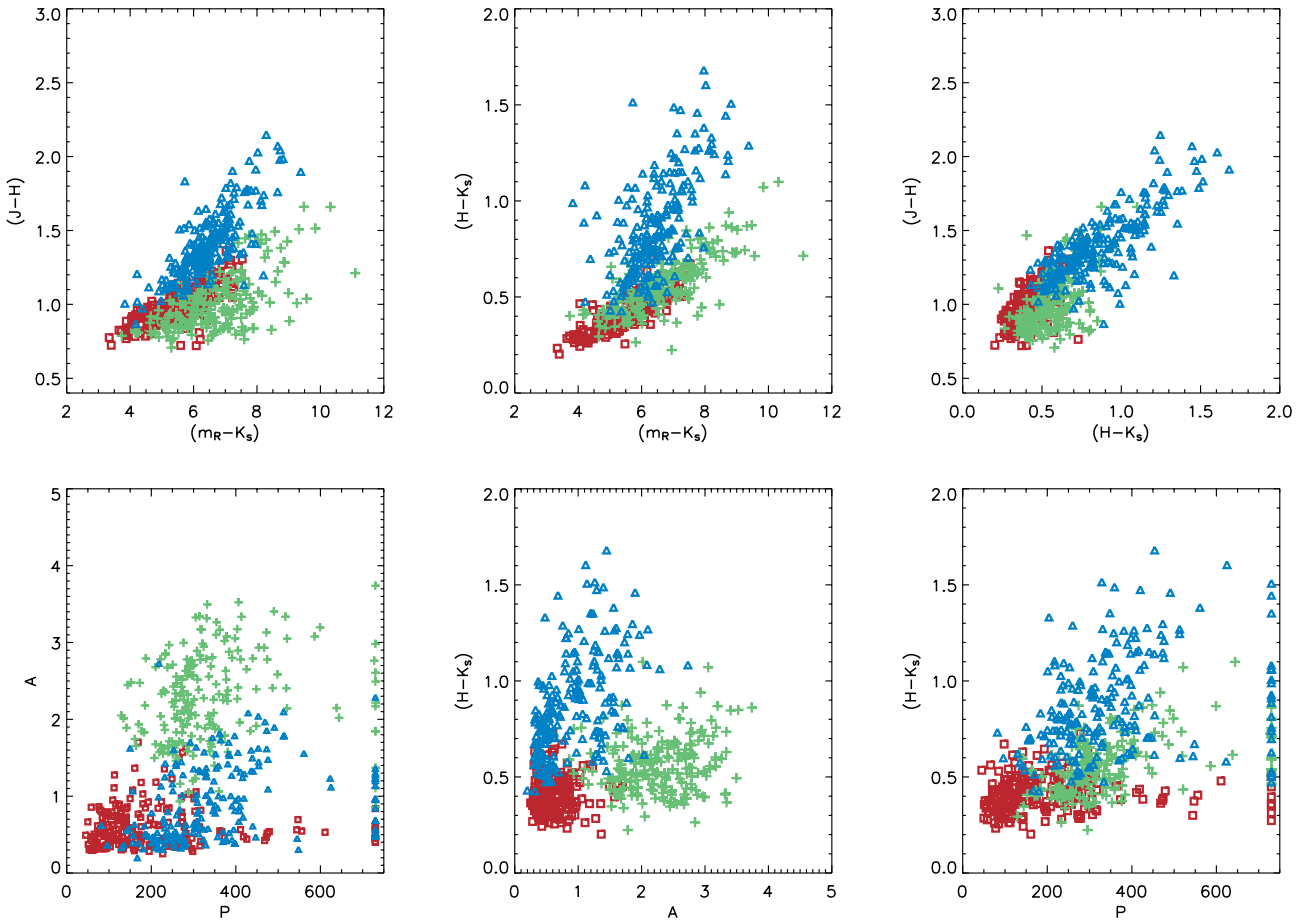

FIG. 9.—All catalog stars in various projections of the input feature space for a three-class problem. Symbols are the same as in Fig. 8. For clarity, only up to 200 objects per class are shown.

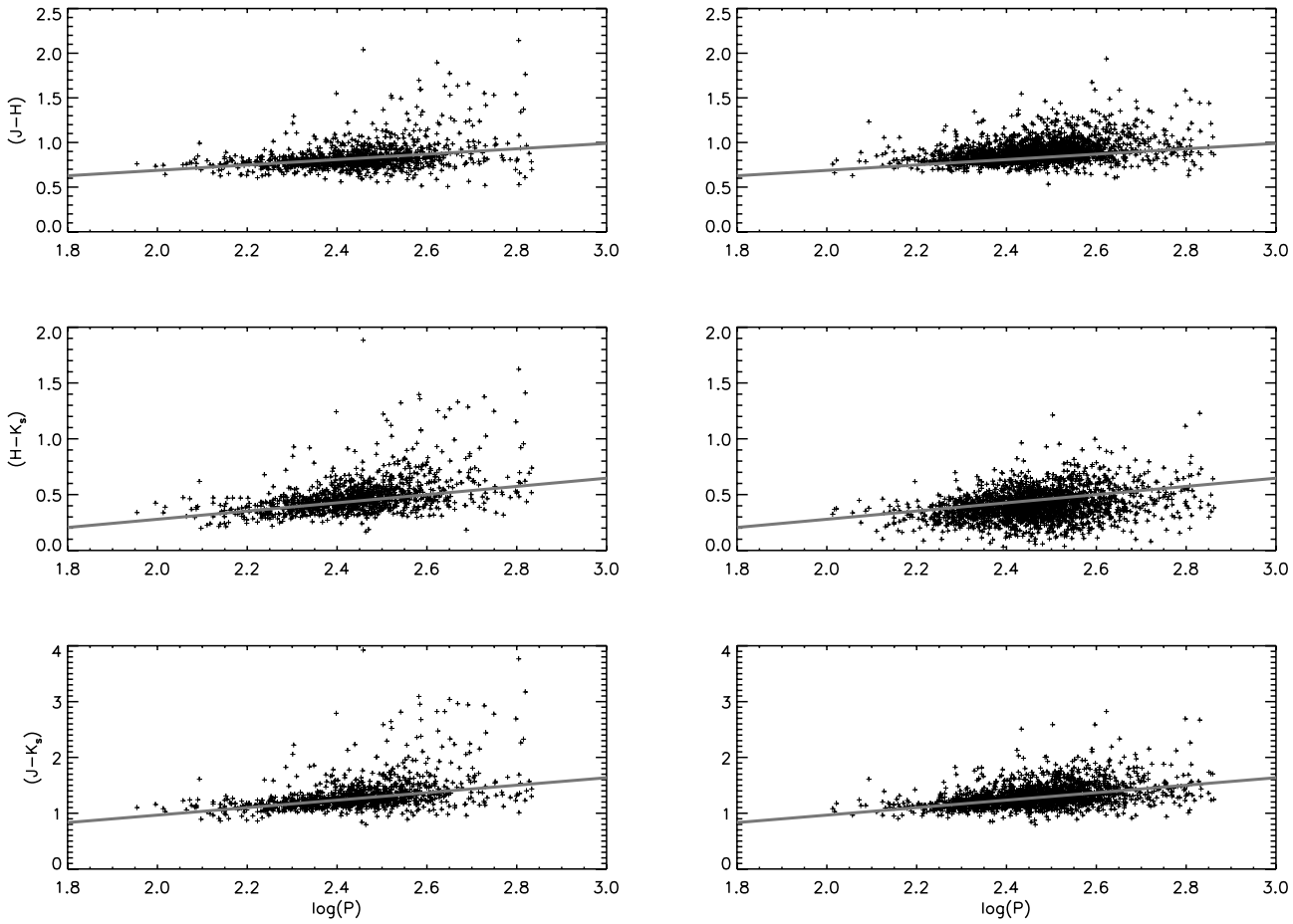

FIG. 10.—Dereddened near-infrared period-color diagrams for Mira variables in the training set (*left*) and among all catalog stars (*right*). Note that the lines are not fits to the plotted data. The lines are fits of Whitelock et al. (2000) to the locus of their Mira and Mira-like semiregular variables when transformed to the 2MASS filter system.

of the input space are displayed in Figures 8 and 9 for both the training and the full classified data sets. As expected, the amplitudes of carbon stars are lower than those of Mira-like variables at the same period (Mattei et al. 1997; Mennessier et al. 1997; Whitelock et al. 2000). These objects tend to populate the amplitude gap in Figure 4 and confuse the separation of Miras from other stars. Separating them in color dimensions makes for a better defined gap in the period-amplitude diagram and allows better discrimination between the two remaining classes. However, it should be stressed that, while statistically our class C is dominated by carbon stars, each individual case will require a spectrum for actual confirmation.

5. PERIOD-COLOR RELATIONS

5.1. *Near-IR Colors*

The near-infrared *JHK* colors of red AGB variables have been studied by Whitelock et al. (2000) for a sample of 193 stars detected by *Hipparcos*. The ROTSE-I unfiltered optical band is similar to the *Hipparcos* H_P band, making comparison of the two samples more straightforward. In Figure 10 we plot three dereddened near-infrared colors that can be formed out of 2MASS photometry as a function of $\log P$ for Mira subsamples in our training group (1221 out of 2095 stars) and the final classified data set (2565 out of 8678 stars). The lines are fits from Whitelock et al. to the locus of their Mira and Mira-like semiregular variables, when transformed to the 2MASS filter system. The agreement between their sample and ours is very good, confirming that objects of our class M, selected by the SVM classifier, are indeed mostly Mira variables.

5.2. *Broadband Visual versus Near-Infrared Color*

By considering photometric and kinematic information for a sample of about 350 oxygen-rich M and SR variables, Barthés et al. (1999) were able to distinguish four groups of LPVs. The relative locus of their Mira-dominated groups 1 and 4 versus SR-dominated groups 2 and 3 in the period-color diagram using $V - K$ color is largely reproduced in Figure 11.

One of the findings in Whitelock et al. (2000) was a secondary sequence of Miras and Mira-like SRs with periods shorter than $\log P = 2.35$. These stars, referred to as SP red (short-period red), had broadband $H_P - K$ colors larger by about 1.5 mag than SP blue stars at the same period. The SP blue sequence was more consistent with the continuation of the single sequence at $\log P > 2.35$. We checked for the presence of SP red and SP blue sequences in our data. The left panel of Figure 11 shows our broadband $m_R - K_s$ color plotted against $\log P$ for classes M, SR+L, and C. The line $m_R - K_s = 11.7 \log P - 19.3$ divides the two sequences in Whitelock et al. data and is also shown in Figure 11 (*left*). The general slope of our M sequence is consistent with that of Whitelock et al. and the line. The fact that the boundary is quite far from our M stars most likely comes from the slightly redder range of the ROTSE bandpass. We find no evidence of bimodality within the group classified as "Mira"

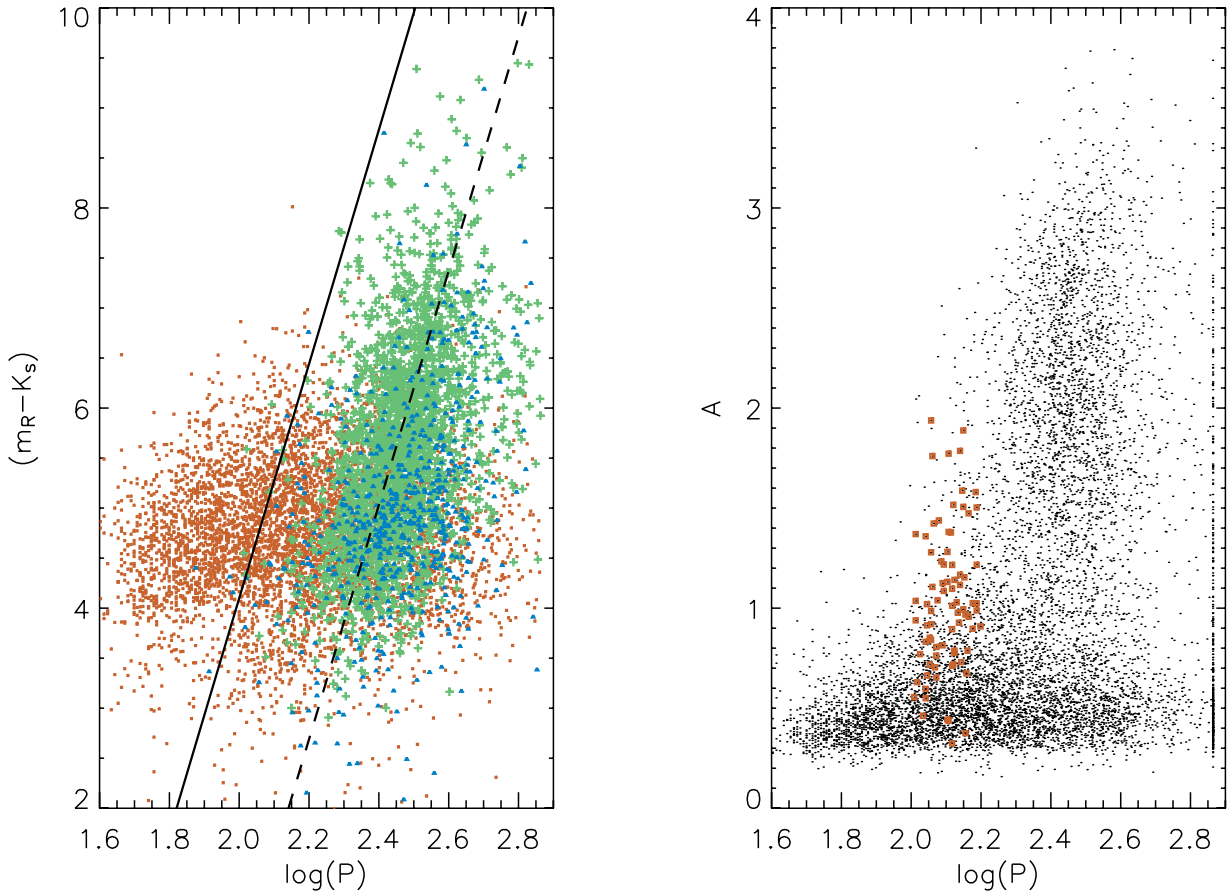

FIG. 11.—Period-color diagram for the broadband optical-infrared colors of all catalog stars (*left*). Symbols are the same as in Figs. 7 and 8. The solid line is the boundary between SP red and SP blue sequences of Whitelock et al. (2000). We note a small overdensity of points near $\log P = 2.1$ and $m_R - K_s = 3.25$. The amplitudes of stars near that locus are among the largest in the SR+L group at the same period (*right*).

at any period range. However, the spread of the primary sequence of M stars is quite large because of observational errors and non-simultaneity of the color data. Another factor affecting our discriminating power is that SR variables, commonly having shorter periods, are assigned to class SR+L ("other") together with L stars. We note, however, the presence of a slight overdensity of SR+L stars with $2.0 < \log P < 2.2$ and $3.0 < m_R - K_s < 3.5$. Out of 79 stars in that region, 24 have GCVS classification with 17 classified as SR. Amplitudes of those objects tend to be in the upper range of the SR+L class or are really more typical of the M class. If the SP red sequence of Whitelock et al. (2000) is actually dominated by Mira-like semiregulars, it will be assigned to the SR+L class in our data, as the majority of SRs reside below the amplitude gap. Therefore, the overdensity near $\log P = 2.1$ and $m_R - K_s = 3.25$ could be, in principle, the short-period end of the SP red sequence, but the evidence is marginal.

6. COMPLETENESS AND RELIABILITY

6.1. *Recovery Rate of LPVs from GCVS*

A rough estimate of the completeness for our catalog can be obtained from the recovery rates of GCVS variables of types M, SR, and L. Deriving the actual completeness requires a detailed simulation rather than comparison with GCVS—which is subject to its own biases—and is left for a future study. The GCVS recovery rates are shown in Figure 12. Note that the recovery rate is only up to 40% in the optimal magnitude range for a typical field. This is most likely a result of our strong requirements on number of good photometric points in the light curve, combined with severe blending near the Galactic plane where red variables concentrate. The rates are still comparable to those of Akerlof et al. (2000), who used much more relaxed selection criteria but only 4 months of data. Therefore, many more lower confidence candidate LPVs with fewer measurements are likely present in the NSVS. Fewer red variables are selected at lower declinations, where the number of available

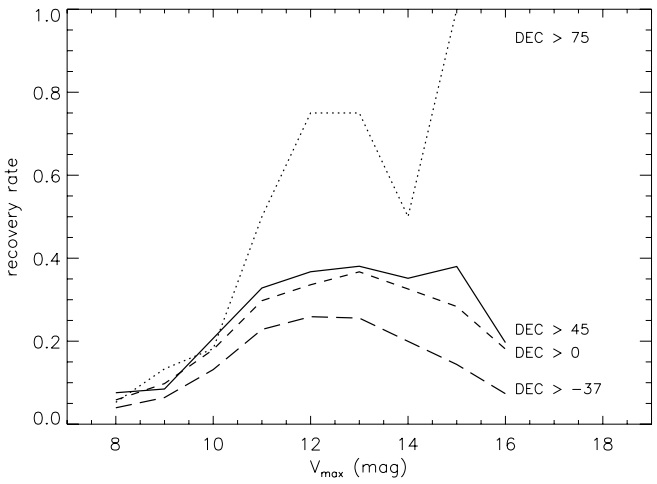

FIG. 12.—Recovery rate for GCVS stars of types M, SR, and L as a function of the V magnitude at maximum light in several declination ranges.

TABLE 4
CATALOG OF RED VARIABLES IN NSVS

NSVS DATA								GCVS DATA		2MASS DATA			
ID	α_{2000}	δ_{2000}	m_R	P	A	Type[a]	Type[b]	Name	Type	ID	K_s	$(J-H)$	$(H-K_s)$
0639385+071951	99.91023	7.33072	11.615	301	1.074	SR+L	SR+L	V0497_Mon	M	06393824+0719524	5.510	0.949	0.567
1553588−183017	238.49516	−18.50478	10.909	319	1.423	M	M	DS_Lib	M	15535891−1830167	5.292	0.918	0.569
1926383+181555	291.65962	18.26522	12.923	730	0.796	M	C	...	...	19263792+1815596	4.517	1.874	1.177
1943437+193408	295.93210	19.56879	10.160	376	0.823	SR+L	C	...	...	19434386+1934081	2.902	1.492	1.434
2002585+410510	300.74363	41.08613	10.830	237	0.408	SR+L	SR+L	...	...	20025852+4105103	4.216	0.945	0.541

NOTE.—Table 4 is presented in its entirety in the electronic edition of the Astronomical Journal. Units of right ascension, and declination are degrees.
[a] Two-class problem: M, SR+L.
[b] Three-class problem: M, SR+L, C.

measurements in the NSVS is diminishing. The narrow magnitude range of the NSVS survey due to heavily sky-dominated photon statistics is responsible for the quick falloff of the efficiency on both ends of the optimal region. At magnitude 9.5, most stars are already saturated most of the time. A recovery rate as high as 75% is possible, but it is limited to a small number of the near-polar objects accessible throughout the year.

6.2. *Possible Misclassifications*

The simulated matches of points to 2MASS sources and selection of color used in § 3.3 do not account for a possible presence of strongly variable, sufficiently red objects of types other than M, SR, and L. Although the variables we are trying to find are by far the most common in the selected part of the parameter space, it is interesting to see what kinds of contaminants we may expect. For this purpose, we correlated the positions of our variables with all GCVS variables. There were 78 matched stars with GCVS classifications of types other than M, SR, and L. Assuming that the relative representation of variability types in GCVS is not vastly different from the actual one, we expect about 3.5% contamination of our sample by stars of other types. It is interesting that six of those stars were classified in GCVS as RR Lyrae with unknown periods and seven were listed as eclipsing binaries. Our light curves of those objects definitely look more like LPV light curves than those of RR Lyraes or E stars. Severe blending might be the explanation in some cases. Several stars were of poorly known irregular types, which, with additional information, could be classified as L, or fairly exotic objects such as progenitors of slow novae that often are in fact of types SR and M (Kholopov et al. 1998). RV Tau stars are related to SRd variables (Percy & Kolin 2000), which were only partially rejected with the cuts discussed in § 3.3. There were 12 RV Tau objects in the contaminating group. Finally, the single largest group was composed of IS-type rapid irregulars (16) and IN Ori variables (10) related to young systems of T Tau type and distinguished based on the presence of the nebula. The full list of GCVS counterparts to our objects is available with the catalog.

7. CATALOG AND DATA ACCESS

Table 4 presents relevant information from the NSVS, GCVS, and 2MASS surveys for selected catalog objects. All data presented in this paper are also available from the SkyDOT database,[5] both as the integral part of that database and in the form of stand-alone text files. There are five files with one line of aggregate information per star and 9371 files with individual light curves. This includes partial data for 693 stars that did not match 2MASS or did not satisfy color cuts. The distribution contains original and derived NSVS catalog information, combined positions and color information from NSVS and 2MASS, input features and results from two classification solutions (§ 4), and, finally, GCVS entries for relevant stars. Light-curve files list measurements from all fields where a given object was independently detected. Measurements are tagged with the NSVS object identification from the SkyDOT database. The data set presented in this paper will also become available through the CDS archive in Strasbourg. Quality flags are explained in Woźniak et al. (2004). For detailed column headers, see the help files available with the distribution.

8. SUMMARY

We searched for well-observed red variables in the NSVS monitoring data and found 8678 objects variable on timescales of tens of days or longer with a red 2MASS counterpart within 1 ROTSE-I pixel (14.″4).[6]

We classified catalog objects using support vector machines, a powerful new machine-learning method publicly available in standard implementations. The present coverage and available features allow good separation of Mira variables from other types of LPVs but are generally insufficient to separate types SR and L. The main feature distinguishing Mira variables from other kinds of red variables is their large amplitude for any given period. We also identified a distinct group of very red stars that are best interpreted as carbon stars. Color information is most important for identifying carbon stars. Based on our classification of 8678 stars into three types, 2276 are Mira variables, 5719 are SR or L variables, and 683 are likely carbon stars. These classes include an estimated ∼1050, ∼4800, and ∼600 new identifications, respectively.

We find no evidence for the SP red and SP blue separation within the class of Mira variables with $\log P < 2.3$ that was described by Whitelock et al. (2000). The effect may be masked by noise, filter differences, or the specifics of object selection. There is marginal evidence that SP red stars in our data set have been assigned SR+L classification and are responsible for the feature near $\log P = 2.1$, $m_R - K_s = 3.25$ in the period-color diagram (Fig. 11).

But the period-amplitude and all period-color diagrams for our Mira class closely resemble those from the previous work

[5] At http://skydot.lanl.gov.

[6] The corresponding catalog of broadband optical light curves, along with the 2MASS colors, is available from the SkyDOT database at http://skydot.lanl.gov.

of Whitelock et al. (2000), confirming that the SVM algorithm is able to distinguish physical classes. While the final automatic classification matches the visual impression of how the input feature space should be partitioned, the machine-learning solution is certainly more objective than visual inspection and incorporates information from all five dimensions in a statistically rigorous way.

The adoption of new data-mining techniques will become even more important as increasingly detailed observations are made and more complex features are included. The SVM technique can handle very complex inputs, and using complete medium-resolution spectra as feature vectors is entirely feasible. Some of the limitations of this study, e.g., the inability to examine cycle-to-cycle variations of light-curve morphology, would be removed if light curves with longer time baselines were available. This underscores the importance of continuous long-term sky monitoring.

This work was supported by Department of Energy contract W-7405-ENG-36 to the RAPTOR project. P. W. acknowledges the Oppenheimer Fellowship at Los Alamos National Laboratory.

REFERENCES

Akerlof, C., et al. 2000, AJ, 119, 1901
Banday, A. J., Zaroubi, S., & Bartelmann, M., ed. 2001, Mining the Sky (Heidelberg: Springer)
Barthés, D., Luri, X., Alvarez, R., & Mennessier, M. O. 1999, A&AS, 140, 55
Bessell, M. S., & Brett, J. M. 1988, PASP, 100, 1134
Chang, C.-C., & Lin, C.-J. 2001, LIBSVM: A Library for Support Vector Machines (Taipei: Natl. Taiwan Univ.), http://www.csie.ntu.edu.tw/~cjlin/libsvm
Cioni, M. R. L., Marquette, J. B., Loup, C., Azzopardi, M., Habbing, H. J., Lasserre, T., & Lesquoy, E. 2001, A&A, 377, 945
Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge: Cambridge Univ. Press)
Eggen, O. J. 1975, ApJ, 195, 661
Feast, M. 2004, in IAU Colloq. 193, Variable Stars in the Local Group, ed. D.W. Kurtz & K. Pollard (ASP Conf. Ser. 310) (San Francisco: ASP), 304
Feast, M., & Whitelock, P. 2000, MNRAS, 317, 460
Feast, M. W., Whitelock, P. A., & Carter, B. S. 1990, MNRAS, 247, 227
Humphreys, R. M., Karypis, G., Hasan, M., Kriessler, J., & Odewahn, S. C. 2001, BAAS, 33, 1322
Ita, Y., et al. 2002, MNRAS, 337, L31
Keeley, D. A. 1970, ApJ, 161, 657
Kharchenko, N., Kilpio, E., Malkov, O., & Schilbach, E. 2002, A&A, 384, 925
Kholopov, P. N., et al. 1998, General Catalogue of Variable Stars (4th ed.; Moscow: Nauka)
Luri, X., Mennessier, M. O., Torra, J., & Figueras, F. 1996, A&A, 314, 807
Maffei, P., & Tosti, G. 1995, AJ, 109, 2652
Mattei, J. A., Foster, G., Hurwitz, L. A., Malatesta, K. H., Willson, L. A., & Mennessier, M. O. 1997, in *Hipparcos*—Venice '97, ed. B. Battrick, M. A. C. Perryman, & P. L. Bernacca (ESA SP-402) (Noordwijk: ESA), 269
Mennessier, M. O., Boughaleb, H., & Mattei, J. A. 1997, A&AS, 124, 143
Paczyński, B. 2000, in ASP Conf. Ser. 203, The Impact of Large-Scale Surveys on Pulsating Star Research, ed. L. Szabados & D. Kurtz (San Francisco: ASP), 9
Percy, J. R., & Kolin, D. L. 2000, J. AAVSO, 28, 1
Pojmański, G. 2002, Acta Astron., 52, 397
Reid, M. J., & Goldston, J. E. 2002, ApJ, 568, 931
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schwarzenberg-Czerny, A. 1989, MNRAS, 241, 153
Vapnik, V. 1998, Statistical Learning Theory (New York: Wiley)
Whitelock, P., Marang, F., & Feast, M. 2000, MNRAS, 319, 728
Willson, L. A. 2000, ARA&A, 38, 573
Wood, P., et al. 1999, in IAU Symp. 191, Asymptotic Giant Branch Stars, ed. T. le Bertre, A. L. Lebre, & C. Waelkens (San Francisco: ASP), 151
Woźniak, P. R., et al. 2004, AJ, 127, 2436
———. 2001, BAAS, 33, 1495
Wray, J. J., Eyer, L., & Paczyński, B. 2003, MNRAS, 349, 1059
Zhang, Y., & Zhao, Y. 2003, PASP, 115, 1006