The Astrophysical Journal, 793:23 (16pp), 2014 September 20 doi:10.1088/0004-637X/793/1/23

SUPERVISED DETECTION OF ANOMALOUS LIGHT CURVES IN MASSIVE ASTRONOMICAL CATALOGS

Isadora Nun[1], Karim Pichara[1,2], Pavlos Protopapas[3,4], and Dae-Won Kim[5]
[1] Computer Science Department, Pontificia Universidad Católica de Chile, Santiago, Chile
[2] The Millennium Institute of Astrophysics, Chile
[3] Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA
[4] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
[5] Max-Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany

Received 2014 March 23; accepted 2014 July 3; published 2014 September 2

ABSTRACT

The development of synoptic sky surveys has led to a massive amount of data for which resources needed for analysis are beyond human capabilities. In order to process this information and to extract all possible knowledge, machine learning techniques become necessary. Here we present a new methodology to automatically discover unknown variable objects in large astronomical catalogs. With the aim of taking full advantage of all information we have about known objects, our method is based on a supervised algorithm. In particular, we train a random forest classifier using known variability classes of objects and obtain votes for each of the objects in the training set. We then model this voting distribution with a Bayesian network and obtain the joint voting distribution among the training objects. Consequently, an unknown object is considered as an outlier insofar it has a low joint probability. By leaving out one of the classes on the training set, we perform a validity test and show that when the random forest classifier attempts to classify unknown light curves (the class left out), it votes with an unusual distribution among the classes. This rare voting is detected by the Bayesian network and expressed as a low joint probability. Our method is suitable for exploring massive data sets given that the training process is performed offline. We tested our algorithm on 20 million light curves from the MACHO catalog and generated a list of anomalous candidates. After analysis, we divided the candidates into two main classes of outliers: artifacts and intrinsic outliers. Artifacts were principally due to air mass variation, seasonal variation, bad calibration, or instrumental errors and were consequently removed from our outlier list and added to the training set. After retraining, we selected about 4000 objects, which we passed to a post-analysis stage by performing a cross-match with all publicly available catalogs. Within these candidates we identified certain known but rare objects such as eclipsing Cepheids, blue variables, cataclysmic variables, and X-ray sources. For some outliers there was no additional information. Among them we identified three unknown variability types and a few individual outliers that will be followed up in order to perform a deeper analysis.

Key words: catalogs – methods: data analysis – methods: statistical – stars: statistics – stars: variables: general

Online-only material: color figures

1. INTRODUCTION

Several important discoveries in astronomy have occurred serendipitously while astronomers were examining other effects. For example, William Herschel discovered Uranus on 1781 March 13 (Herschel 1857) while surveying bright stars and nearby faint stars. Similarly, Giuseppe Piazzi found the first asteroid, Ceres, on 1801 January 1 (Serio et al. 2002) while compiling a catalog of star positions. Equally unexpected was the discovery of the cosmic microwave background (CMB) radiation in 1965 by Arno Penzias and Robert Wilson while testing the Bell Labshorn antenna (Penzias & Wilson 1965).

With the proliferation of data in astronomy and the introduction of automatic methods for classification and characterization, the keen astronomer has been progressively removed from the analysis. Anomalous objects or mechanisms that do not fit the norm are now expected to be discovered systematically: serendipity is now a machine learning task. As a consequence, the astronomer's job is no longer to be behind the telescope, but to be capable of selecting and interpreting the increasing amount of data that is provided by technology.

Outlier detection, as presented here, can guide the scientist to identify unusual, rare, or unknown types of astronomical objects or phenomena (e.g., high-redshift quasars, brown dwarfs, pulsars). These discoveries might be useful not only to provide new information, but also to outline observations, which might require further and deeper investigation. In particular, our research detects anomalies in photometric time series data (light curves). For this work, each light curve is described by 13 variability characteristics (period, amplitude, color, etc.) called *features* (Kim et al. 2011; Pichara et al. 2012), which have been used for classification. It is worth noting that the method developed in this paper is not only applicable to time-series data, but could also be used for any type of data that need to be inspected for anomalies. In addition to this advantage, the fact that it can be applied to large data makes this algorithm suitable for almost any outlier detection problem.

Many outlier detection methods have been proposed in astronomy. Most of them are unsupervised techniques, where the assumption is made that there is no information about the set of light curves or their types (Xiong et al. 2010). One of these approaches considers a point-by-point comparison of every pair of light curves in the database by using correlation coefficient (Protopapas et al. 2006). Other techniques search for anomalies in lower-dimensional subspaces of the data in order to deal with the massive number of objects or the large quantity of features that describe them (Henrion et al. 2013; Xiong et al. 2010). Clustering methods are equally applied in the astronomical

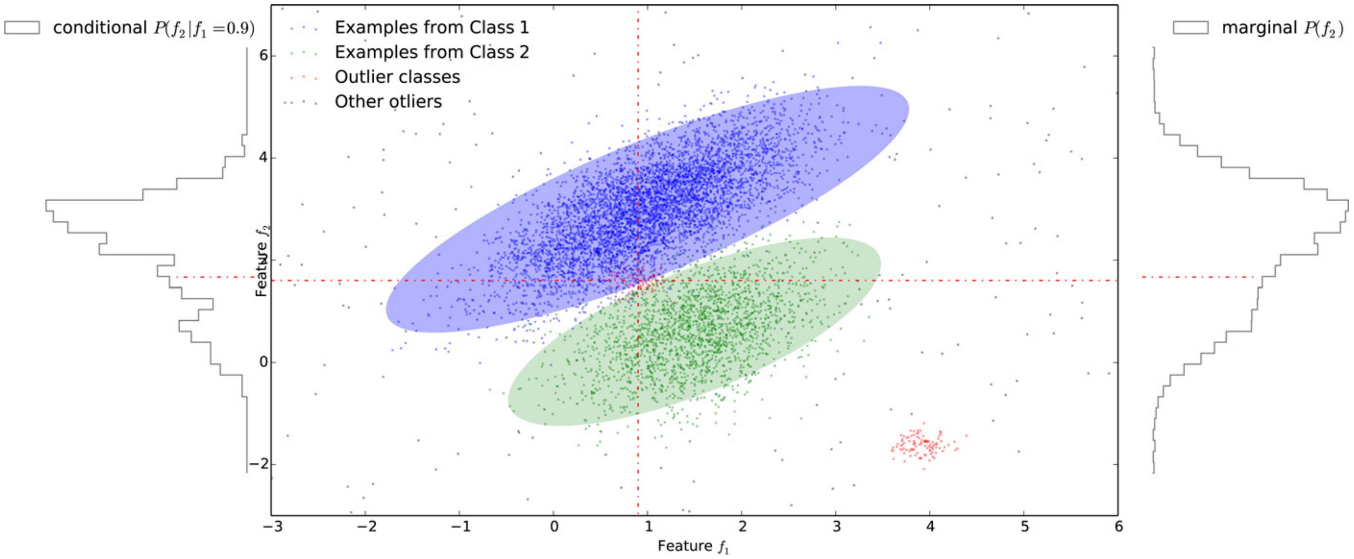


Figure 1. Simple illustration of the method. In most unsupervised methods the red points in the middle will not be considered as outliers because they are in a region with point density that is not separable. The product of the probabilities or the sum of the distances to the known classes may not be adequate as an outlier score, and therefore the joint probability is a better measure for outliers. This case occurs when the conditional probability is lower than the marginal probability, as can be seen from this simple illustration.

(A color version of this figure is available in the online journal.)

outlier detection area aiming to find clusters of new variability classes (Bhattacharyya et al. 2012; Rebbapragada et al. 2008). Unfortunately, these methods either scale poorly with massive data sets and with high-dimensional spaces or partially explore the data, therefore missing possible outliers.

In this paper, we face these constraints by creating an algorithm that is able to efficiently deal with large data and is capable of exploring the data space as exhaustively as possible. Furthermore, we address this matter from a point of view different from that presented by He & Carbonell (2006) as the new-class discovery challenge.Contrary to unsupervised methods, it relies on using labeled examples for each known class in the training set, and, unlike supervised methods, we assume the existence of some rare classes in the data set for which we do not have any labeled examples. This approach takes advantage of available information, but it does not restrict the anomalous findings to a certain type of light curves. Furthermore, in unsupervised anomaly detection methods, in which no prior information is available about the abnormalities in the data, anything that differs from the entire data set is flagged as an outlier, and, consequently, many of the anomalies found would simply be noise. In contrast to these techniques, supervised methods incorporate specific knowledge into the outlier analysis process, thus obtaining more meaningful anomalies. This is illustrated in Figure 1. The blue and green points represent instances in a two-dimensional feature space from the known class 1 and class 2, respectively. The shaded areas represent the boundaries learned from a classifier. The gray points represent isolated outliers, and the red points represent outlier classes. In most unsupervised methods, the red points in the middle will not be considered as outliers because they are in a region with point density that is not separable. In the most naive supervised methods, anything that is outside the boundaries is considered as an outlier. For the example of the outlier class in the middle, the product of the probabilities or the sum of the distances to the known classes may not be adequate as an outlier score, and therefore the joint probability is a better measure for outliers. This case occurs when the conditional probability is lower than the marginal probability,[6] as can be seen from this simple illustration. The conditional probability shown on the left is smaller than the marginal probability shown on the right. Our model will consider those objects as outliers.

In the first stage of our method, we build a classifier that is trained with known classes (every known object is represented by its features and a label). We then use the classifier decision mechanism to our advantage. More precisely, we learn a probability distribution for the classifier votes on the training set in order to model the behavior of the classifier when the objects correspond to a known variability class. The intuition behind this method is to recognize, and thus to learn, how the classifier is confused when it comes to voting. By confusion, we refer not only to the hesitation between two or more classes for an object label but also to the weights it assigns to each of these possibilities.

Therefore, when an unlabeled light curve is fed into the model, the classifier attempts to label it, and if this classifying behavior is known by the model, the object will have a high probability of occurrence and consequently a low outlierness score. On the contrary, the object will have a higher anomaly score and will be flagged as an outlier candidate insofar as the classifier operates in a different manner from the previously known mechanisms.

Once our outlier candidates are selected, an iterative post-analysis stage becomes necessary. By visual inspection we discriminate artifacts from true anomalies, and we (1) remove them systematically from our data set and (2) create classes of spurious objects that we add to our training set. We then rerun the algorithm and obtain new candidates. These steps are repeated until we do not obtain any apparent artifacts in our outlier list and a clustering method is finally executed. The purpose of this phase is to group similar objects in new variability classes and to consequently give them an astronomical interpretation. Finally, we cross-match the most interesting outliers with all publicly available catalogs with the aim of verifying whether

[6] This is not necessarily true for all cases.

there is any additional information about them. In particular, we are interested in knowing whether they belong to a known class. In the negative case, the outliers will be followed up using spectroscopy to deeply analyze their identity and behavior.

To achieve this, we use random forest (RF; Breiman 2001) for the supervised classification in order to obtain the labeling mechanism for each class on the training set. RF has been extensively and successfully used in astronomy for cataloging (Pichara & Protopapas 2013; Kim et al. 2014). Starting with the RF output, we construct a Bayesian network (BN) with the purpose of extracting the classification patterns that we use for our final score of outlier detection.

The paper is organized in the following way. Section 2 is devoted to other methods related to anomaly detections in machine learning and astronomy. In Section 3 we detail the background theory, including the basic blocks of an RF network and a BN. Our approach and the pipeline followed in the paper are shown in Section 4. Section 5 contains information about the data used in this work, and Section 6 presents the results of the tests performed and the experiments with real data, including retraining and eliminating artifacts. We proceed by explaining in Section 7 the post-analysis process. Finally, conclusions follow in Section 8.

2. RELATED WORK

2.1. Outlier Detection in Machine Learning

A vast amount of literature has been published in relation to anomaly/outlier detection problems (Chandola et al. 2009; Kou et al. 2004), but it can generally be classified into two main classes: supervised and unsupervised methods.

In unsupervised approaches, the examples given are unlabeled, and consequently there is no training set in which the data are separated into different classes. In turn, these techniques can be partitioned into three main subcategories: statistical methods, proximity-based methods, and clustering methods.

Statistical approaches are the earliest methods used for anomaly detection. These methods detect anomalies as outliers that deviate markedly from the generality of the observations (Grubb & Frank 1969) by assuming that a statistical model generates normal data objects and that data that do not follow the model are outliers. In particular, many of these methods use mixture models by applying Gaussian distributions (Agarwal 2005; Eskin 2000). The typical strategy considers the calculation of a score and a threshold, both used to identify points that deviate from normal data. For example, Eskin (2000) proposes an algorithm that fits mixture models, normal and anomalous, using the expectation maximization (EM) algorithm and assuming a prior probability λ of being anomalous. Then, the author obtains an anomaly score that is based on measuring the variation of the normal distribution when a point is moved to the anomalous distribution. One of the main drawbacks of the statistical approach is that it is generally applied to quantitative data sets or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical processing. This limits its range of application and can increase the processing time when complex data transformations are needed as a preprocess. (Hodge & Austin 2004).

Clustering-based methods (Yang et al. 2006; Son et al. 2009; Zhang et al. 1996) are based on the fact that similar instances can be grouped into clusters and that normal data lie on large and dense clusters, while anomalies belong to small or sparse clusters, or to no cluster at all. Most recent clustering algorithms proposed for anomaly detection are in the context of intrusion detection on networks (Yang et al. 2006; Son et al. 2009). Unfortunately, clustering algorithms suffer from a dimensionality problem. Often, in large-dimensional spaces, distance metrics that are applied to characterize similarity do not provide suitable clusters. Subspace clustering algorithms, a remedy to the dimensionality curse, have not been commonly used for anomaly detection, with the exception of some recent works (Seidl et al. 2009; Pichara et al. 2008; Pichara & Soto 2011). Seidl et al. (2009) perform a subspace clustering algorithm to rank data points according to the size of the clusters and the number of dimensions of each subspace where the points belong. To identify microclusters containing anomalies, Pichara et al. (2008) search for relevant subspaces in subsets of variables that belong to the same factor in a trained BN. Similarly, Pichara & Soto (2011) present a semi-supervised algorithm that actively learns to detect anomalies in relevant subsets of dimensions, where dimensions are selected by using a subspace clustering technique that finds dense regions in a sparse multidimensional data set. One of the main drawbacks of these kinds of approaches is that they use heuristics to find relevant subspaces, and those heuristics may ignore combinations of spaces where anomalies could also lie.

Finally, proximity-based methods follow the intuition that anomalies are records with fewer neighbors than normal records (Ramaswamy et al. 2000; Knorr & Ng 1998; Breunig et al. 2000). For example, Breunig et al. (2000) assign an anomaly score called the local outlier factor (LOF) to each data instance; this score is given by the ratio between the local density of the point and the average local density of its k-nearest neighbors. Local density is calculated using the radius of the smallest hypersphere that is centered at the data instance and contains k (nearest) neighbors. Papadimitriou et al. (2003) propose a variant of the LOF called the multi-granularity deviation factor (MDEF). For a given record, its MDEF is calculated as the standard deviation among its local density and the local densities of its k-nearest neighbors. Then they use the MDEFs to search microclusters of anomalous records. Along the same lines, Jin et al. (2001) propose another variant of LOF that improves efficiency by avoiding unnecessary calculations. They achieve this by calculating upper and lower bounds among the microclusters detected. Unfortunately, density-based algorithms are usually quadratic in the number of instances, and thus they are not suitable for large data. Furthermore, these methods also suffer from the curse of dimensionality for the same reasons mentioned above for the clustering methods.

On the other hand, in supervised approaches, outlier detection can be treated as a classification problem, where a training set with class labels is used to generate a classifier that distinguishes between normal and anomalous data (Gibbons & Matias 1998; Aggarwal & Yu 2001; Chandola et al. 2009). Various anomaly detection algorithms have been proposed in this area, such as decision trees (John 1995; Arning et al. 1996) and neural networks (Nairac et al. 1999; Bishop 1994). Decision tree algorithms fit the data focusing only on salient attributes, a desirable characteristic when dealing with high-dimensional data. These algorithms work by modeling all points corresponding to normal classes; then points having an erroneous or unexpected classification are considered as anomalies. Similarly, neural networks are employed to model the unknown distribution of normal class points by training a feed-forward network. This is achieved by adjusting the weights and thresholds while learning from the input data. Neural networks work well when training sets are representative of the unseen data. Unfortunately, this may not occur for new instances that are beyond of the scope of the training set. Decision trees and neural networks are susceptible to overfitting when stopping criteria are not well determined.

2.2. *Outlier Detection in Astronomy*

Because synoptic sky surveys have significantly increased in the past decade (Keller et al. 2007; Hodapp et al. 2004; Tyson et al. 2002), astronomical anomaly detection has not yet been fully implemented in the enormous amount of data that has been gathered. As a matter of fact, barring a few exceptions, most of the previous studies can be divided into only two different trends: clustering and subspace analysis methods.

In Rebbapragada et al. (2008), the authors create an algorithm called periodic curve anomaly detection (PCAD), an unsupervised outlier detection method for sets of unsynchronized periodic time series, by modifying the k-means clustering algorithm. The method samples the data and generates a set of representative light curve centroids from which the anomaly score is calculated. In order to solve the phasing issue, during each iteration, every time series is rephased to its closest centroid before recalculating the new one. The anomaly score is then calculated as the distance of the time series to its closest centroid. Even if the anomaly detection is satisfactory on a restricted and small data set, the technique scales poorly with massive data sets. This is mainly due to the distinctive high-dimensionality problem that clustering methods encounter, as mentioned in the previous section. Furthermore, since the algorithm is based on the alignment of the time series periods, it is restricted to periodic light curves, thus limiting the scope of possible astronomy applications.

Similarly, Protopapas et al. (2006) search for outlier light curves in catalogs of periodic variable stars. To this end, they use cross-correlation as a measure of similarity between two individual light curves and then classify light curves with the lowest average similarity as outliers. Unfortunately, this method scales as N_{LC}^2, where N_{LC} is the number of light curves. In order to deal with this high operational cost and to apply the algorithm to large data sets, they make an approximation they call *universal phasing*. By using clustering, they find where the signal with the highest/lowest magnitude dip occurs for each light curve and set it to a particular phase by time-shifting the folded light curve. Once they find an absolute phase for all the light curves, they calculate the correlation of each one with the average of the rest of the set, reducing the operational cost of the algorithm to N_{LC}. Unfortunately, this method is an approximation since it does not guarantee that the correlation between two light curves is maximum. Furthermore, this approximation also implies not taking into account the observational errors, thus losing highly valuable information. Finally, as in Rebbapragada et al. (2008), this algorithm is also restricted to periodic light curves.

Xiong et al. (2010) separate astronomy anomalies into two different categories: *point anomalies*, which include individual anomalous objects, such as single stars or galaxies that present unique characteristics, and *group anomalies* (anomalous groups of objects), such as unusual clusters of the galaxies that are close together. For that end, they develop one method for each of these cases. For the former case, the authors create mixed-error matrix factorization (MEMF), an unsupervised algorithm that explores subspaces of the data. They also assume that normal data lie in a low-dimensional subspace and that their features can be reconstructed by linear combination of a few base features. Quite the opposite, anomalies lie outside of that subspace and cannot be well reconstructed by these bases. To do so, they find a robust low-rank factorization of the data matrix and consider the low-rank approximation error to be an additive mixture of the regular Gaussian noise and the outliers that can be measured differently in the model. One limitation of MEMF is that the factorization rank k has to be specified by the user and is consequently often determined by heuristics. For group anomalies, the authors use hierarchical probabilistic models to capture the generative mechanism of the data. In particular, they propose the Dirichlet genre model (DGM), which assumes that the distribution of the groups in the data set can be represented by a Dirichlet distribution. Two anomaly scores are then presented: the likelihood of the whole group and a scoring function that focuses on the distribution of objects in the group. One of the main drawbacks of this method is that the inference stage considers a nonconvex problem and is consequently restricted to the limitation of variational approximations.

Henrion et al. (2013) propose CASOS, an algorithm that detects outliers in data sets obtained by cross-matching astronomical surveys. To do so, they compute an anomaly score for each observation in lower-dimensional subspaces of the data, where subspaces make allusions to subsets of the original data variables. In particular, any anomaly detection method that produces numerical anomaly scores can be used with this approach. The idea is to analyze the anomaly score of each observation in every possible subspace and then combine them in such a way that objects with many observed variables and objects with only a few are equally likely to have high anomaly scores. Unfortunately, CASOS has the disadvantage that it will not be able to detect outliers, which are only apparent in multivariate spaces with significant numbers of variables.

Finally, Richards et al. (2012) apply a semi-supervised approach for astronomical outlier detection. Unlike the previously mentioned algorithms, in this work the authors compute a distance metric from every candidate object to each source in a training set. To do so, they train an RF classifier with known classes and measure the proximity value ρ_{ij} for all the new instances i to every j object on the training set. The proximity measure ρ_{ij} gives the proportion of trees in the RF for which the feature vectors x_i and x_j appear in the same terminal node. Using this proximity measure, they create an outlier score and evaluate each instance in the database. A threshold on the anomaly score is then determined in order to decide whether or not an object is an outlier. This approach suffers from the same constraints as density-based outlier detection methods. It is operationally expensive and slow for large databases since every evaluated object must be compared with each instance on the training set. Furthermore, it has the problem of determining the outlier threshold, in other words, what is considered as a “far” or “close” distance.

3. BACKGROUND THEORY

Our algorithm is based on known machine learning methods, namely, RF and Bayesian networks. In this section, we summarize the background for all these methods. Detailed explanations for each of these approaches can be found in Breiman (2001), Koller & Friedman (2009), and Cooper & Herskovits (1992).

3.1. *Random Forest*

RF developed by Breiman (2001) is a very effective machine learning classification algorithm. The intuition behind this method is to train several decision trees using labeled data (training set) and then use the resulting trained decision trees to classify new unlabeled objects in a voting system. The main principle is to follow a divide-and-conquer approach; each decision tree is trained with a random sample of the data and is consequently considered as a "weak" classifier. Nevertheless, the ensemble of these decision trees generates a robust or "strong" classifier that, based on the combinatorial power of its construction, creates an accurate and effective model.

The process of training or building an RF model given some training data is as follows.

1. Let R be the number of trees in the forest (a user-defined parameter) and $|F|$ be the number of features describing the data.
2. Build R sets (bags) of n samples taken with replacement from the training set (bootstrap samples). Note that each of the R bags has the same number of elements as the training set, but some of the examples are selected more than once, given that the samples are taken with replacement.
3. For each of the R sets, train a decision tree using at each node a feature selected from a random sample of $|F'|$ features ($|F'|$ is a model parameter where $|F'| \ll |F|$) that optimizes the split.

Each decision tree is created independently and randomly using two principles. First, each individual tree is trained on different samples of the training set. Growing trees from different samples of the training set creates the expected diversity among the individual classifiers. The second principle is the random feature selection, which means that for each tree the splitting (decisive) feature in every node is chosen from a random subset of the features. This contributes to the reduction of the dimensionality and has been shown to significantly improve the RF accuracy (Bernard et al. 2008; Geurts et al. 2006). Furthermore, each tree is grown to the maximum possible subject to the minimum size chosen for the terminal nodes (model parameter). For this paper, we set the terminal node minimum size to be 1, so the trees can be as large as the model allows it.

When classifying a new instance, each tree gives a classification or "votes" by following the decision rules in every node of the tree until reaching a terminal node. Since RF is a composition of many trees, the output corresponds to the votes of all the trees. The class probability, $P(\text{class } y_j/\text{features}(i))$, is the proportion of trees that voted for the class j ($j \in \{1, \ldots, k\}$, where k is the total number of classes).

3.2. *Bayesian Networks*

A BN is a directed acyclic graph (DAG), a particular probabilistic graphical model that encodes local statistical dependencies among random variables. A BN is defined by a set of nodes representing random variables $V = \{v_1, \ldots, v_k\}$ and a set of edges $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_b\}$ connecting the variables. One of the applications of a BN is to estimate joint probability density functions (pdfs). This is done by assuming that the variables in the pdf are the nodes in the BN and that the connections between the nodes determine certain dependence relationships that simplify the joint distribution. More formally, if we want to estimate the joint probability distribution $P(v_1, \ldots, v_k)$ and we have a BN describing connections between these variables, we can simplify it as

$$P(v_1, \ldots, v_k) = \prod_{i=1}^{k} P(v_i | Pa_{BN}(v_i)), \tag{1}$$

where $Pa_{BN}(v_i)$ corresponds to the parents of the node v_i in the BN. Note that the pdf has been decomposed in a product of smaller factors (conditional probabilities).

The main challenges of learning a BN that models a PDF over a set of variables are (1) to learn the set of edges ε, or in other words the BN structure, and (2) to learn the conditional probabilities $P(v_i | Pa_{BN}(v_i))$.

3.3. *Learning the Edges of the BN*

A BN is a directed acyclic graph where each node represents a random variable. In our case the random variables we are modeling are the RF outputs, in other words, a probability vector $[v_1, \ldots, v_k]$ $v_j \in [0, 1]$, representing the probabilities of belonging to each of the possible classes c_j ($j \in [1 \ldots k]$, with k being the number of known classes). Given that the amount of possible network structures is exponential in the number of variables, it is necessary to use heuristics to find the optimal network. In our work, we use a greedy algorithm proposed by Cooper & Herskovits (1992). They define a score to evaluate each possible network structure and greedily search for the structure with the maximum score. First, they decide an order of the variables (topological order) from where possible structures will be explored. A topological order $\{1, \ldots, k\}$ is such that if i is smaller than j in the order, then v_i is an ancestor of v_j in the network structure. After deciding on a particular order, the algorithm proceeds by finding the best set of parents per node, greedily adding a new candidate parent and checking whether or not the new addition creates a better network score. In the case in which the edge addition improves the network score, the edge remains in the actual network. Note that the maximum number of parents per node is an input parameter of the algorithm.

Finally, to calculate the network score, they evaluate the probability of the structure given the data, which corresponds to applying the same factorization imposed by the structure to the data and using multinomial distributions over each factor. Exactly how the score is assigned to a given structure is well described in the original work (Cooper & Herskovits 1992; Pichara & Protopapas 2013).

3.4. *Learning the Parameters of the Conditional Distributions*

In order to model the conditional probabilities, we may assume that all variables (votes) are continuous and normally distributed. Since V comes from the RF votes, its distribution is multimodal, and consequently a single Gaussian would not describe the data. A better solution is to discretize the continuous data (Monti & Cooper 1998), so as to use multinomial distributions. Even if this process only gets rough characteristics of the distribution of the continuous variables, it better describes the data by capturing its multimodality. To perform the discretization, the data are divided into a set of bins; thus, every data value that falls in a given interval is replaced by a representative value of that interval.

Given that our data are now discrete, we use multinomial distributions to model each conditional probability $P(v_j | Pa_{BN}(v_j))$. The number of parameters to be estimated depends on the number of values that variables v_j and $Pa_{BN}(v_j)$ can take. For example, suppose that the parents of variable v_j are $\{v_a, v_b\}$, where each of the three variables $\{v_j, v_a, v_b\}$ can take two different values (for simplicity, say, 1 and 2). The probability distribution $P(v_j | v_a, v_b)$ is then completely determined by Table 1.

Table 1
Probability of v_j Given the Different Values of the Parents, $P((v_j|v_a, v_b))$.

	$v_j = 1$	$v_j = 2$
$v_a = 1, v_b = 1$	θ_1	$1 - \theta_1$
$v_a = 1, v_b = 2$	θ_2	$1 - \theta_2$
$v_a = 2, v_b = 1$	θ_3	$1 - \theta_3$
$v_a = 2, v_b = 2$	θ_4	$1 - \theta_4$

Note. There is one multinomial distribution per each combination of the values of the parents. The number of outcomes of each distribution corresponds to the number of values of variable v_j.

The number of parameters for each variable is consequently given by the following expression:

$$(N_{\mathrm{bins}} - 1) \times (N_{\mathrm{bins}})^{N_{\mathrm{parents}}}, \quad (2)$$

where N_{bins} is the number of bins chosen for the discretization and N_{parents} corresponds to the number of parents of the variable. In the example given above, the number of parameters we have to estimate is $(2 - 1) \times 2^2 = 4$. To estimate the parameters, we use the *maximum a posteriori* (MAP) approach, where we select the value for the unknown parameter as the value with maximum probability under the posterior distribution of the parameter. The posterior distribution of, say, θ_1 is calculated as

$$P(\theta_1|data) = \frac{P(data|\theta_1) \times P(\theta_1)}{\sum_{\theta_1} P(data|\theta_1) \times P(\theta_1)}, \quad (3)$$

where $P(data|\theta_1)$ is the likelihood of the model and $P(\theta_1)$ is the prior of the parameter θ_1. The likelihood is calculated as

$$P(data|\theta_1) = \theta_1^{N_1} \times (1 - \theta_1)^{N_2}, \quad (4)$$

where N_1 is the number of cases in the data where v_j takes a particular value. Following the example above, N_1 is the number of cases where $v_j = 1$ and $v_a = 1, v_b = 1$, and N_2 is the number of cases where $v_j = 2$ and $v_a = 1, v_b = 1$.

The main purpose of the priors is to avoid overfitting. In other words, in cases where we have just a few cases in the data with a given combination of values, the estimation of the parameters should tend to stay in a predefined value until the data cases increase. Priors are a way to simulate previously seen "imaginary data" in order to compensate situations of a few cases. We choose conjugate priors to simplify the calculations of the posteriors. In our case, given that the likelihood is a multinomial, the chosen prior for $P(\theta)$ is a Dirichlet distribution, which is the conjugate distribution for the multinomial. Using a Dirichlet prior, the obtained posterior is

$$P(\theta_1|data) \propto \theta_1^{N_1+\alpha_1} \times (1 - \theta_1)^{N_2+\alpha_2}, \quad (5)$$

where $\{\alpha_1, \alpha_2\}$ are the Dirichlet distribution parameters. The values of $\{\alpha_1, \alpha_2\}$ act as the "imaginary data" that we count, and we just assume that all combinations of values have the same number of previously seen cases. Analogously, we can find the value of every parameter θ_j for variables with any number of different values.

4. METHODOLOGY

In the next section, we detail our work and methodology. For illustration, we present in Figure 2 a pictorial representation of our algorithm and its two main stages: the training stage (left panel) and the outlier detection stage (right panel). In the training stage, we start with a training set followed by the training of the RF, discretization of the probabilities, and finally the construction of the BN. In the outlier discovery stage, every new instance is passed through the already-learned RF and BN resulting in a score for being outlier.

Table 2
Training Set Composition

	Class	Number of Objects
1	Nonvariable	3969
2	Quasars	58
3	Be Stars	127
4	Cepheid	78
5	RR Lyrae	288
6	Eclipsing Binaries	193
7	MicroLensing	574
8	Long Period Variable	359

As we previously mentioned, the idea behind our method is to train a classifier with known classes and learn its decision mechanism with a model. In this manner, when an outlier is being analyzed, the classifier will present an abnormal voting confusion that will be immediately flagged by our model.

Our method starts with a set of n labeled instances (training set) $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where each $x_i = \{x_{i1}, \ldots, x_{iD}\}$ is a vector in a D-dimensional space—the statistical descriptors or features that represent each light curve—and y_i corresponding to the label of x_i ($y_i \in \{c_1, \ldots, c_k\}$ are all the known classes in the training set). In Section 5 we give details about the classes and statistical descriptors we used.

We train an RF classifier and obtain voted labels for each element using the set S. Since trees are constructed from different bootstrapped samples of the original data (as explained in Section 3.1), about one-third of the cases are left out of the "bag" and not used in the construction of each tree. By putting these out-of-bag (oob) observations down the trees that were not trained with oob data, we end up with unbiased predicted labels for S. Each prediction obtained from the RF comes as a vector $\{v_{i1}, \ldots, v_{ik}\}$, where each $v_{ij} \in [0, 1]$, $j \in [1, \ldots, k]$ tells us the probability that the element x_i belongs to the class y_j, $\sum_{j=1}^{k} v_{ij} = 1$, $\forall i \in [1, \ldots, n]$, or, as we explained in Section 3.1, v_{ij} corresponds to $P(class\ y_j/features(i))$.

In our experiments we use 20 bins for the discretization and a maximum of two parents.

After this step is performed, we conclude with a new data set $V = \{v_1, \ldots, v_n\}$, where each $v_i = \{v_{i1}, \ldots, v_{ik}\}$. This data set gives us information about how the RF votes among objects that belong to each of the known variability classes. We want to use this data set to decide whether or not an unlabeled object belongs to an unknown variability class, simply by comparing the RF votes of this unlabeled object with the "usual" votes of the RF obtained from the data set V. If the voting vector for the unlabeled object is too different from the voting vectors stored in the data set V, we flag it as an outlier. To do this comparison, we learn the joint probability distribution over the data set V using a BN. Recall that BNs estimate joint probability distributions as a product of smaller factors. These factors are conditional probability distributions, and, in our case, the joint probability we aim to model is the joint probability of the various votes, $P(v_1, \ldots, v_k)$.

When analyzing an unlabeled object i, we first obtain its RF votes $\{v_{i1}, \ldots, v_{ik}\}$ using the already-trained RF, and next

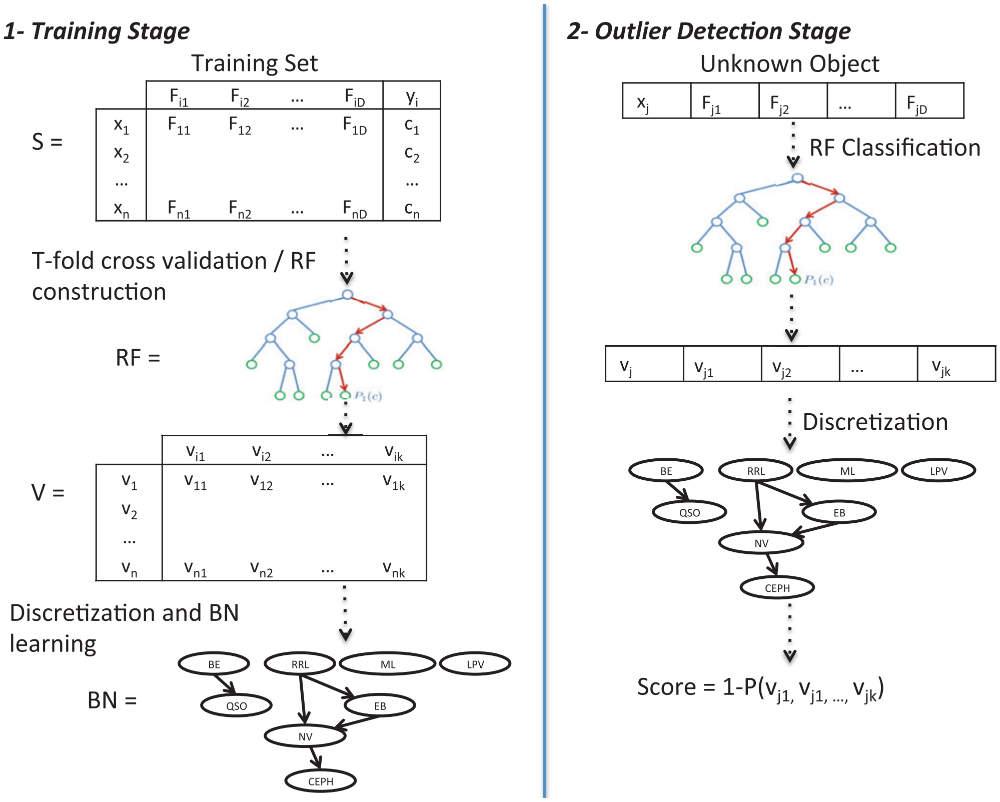

1- Training Stage
Training Set
S =
F$_{i1}$ F$_{i2}$ … F$_{iD}$ y$_i$
x$_1$ F$_{11}$ F$_{12}$ … F$_{1D}$ c$_1$
x$_2$ c$_2$
… …
x$_n$ F$_{n1}$ F$_{n2}$ … F$_{nD}$ c$_n$
T-fold cross validation / RF construction
RF =
$P_i(c)$
V =
v$_{i1}$ v$_{i2}$ … v$_{ik}$
v$_1$ v$_{11}$ v$_{12}$ … v$_{1k}$
v$_2$
…
v$_n$ v$_{n1}$ v$_{n2}$ … v$_{nk}$
Discretization and BN learning
BN =
BE
RRL
ML
LPV
QSO
EB
NV
CEPH
2- Outlier Detection Stage
Unknown Object
x$_j$ F$_{j1}$ F$_{j2}$ … F$_{jD}$
RF Classification
$P_i(c)$
v$_j$ v$_{j1}$ v$_{j2}$ … v$_{jk}$
Discretization
BE
RRL
ML
LPV
QSO
EB
NV
CEPH
Score = 1-P(v$_{j1,}$ v$_{j1, …,}$ v$_{jk}$)

Figure 2. Pictorial representation of the algorithm and its two main stages: the training stage (left panel) and the outlier detection stage (right panel). (A color version of this figure is available in the online journal.)

we calculate the joint probability associated with this vector $P(v_{i1}, \ldots, v_{ik})$ using the already-learned BN. Our outlier score is calculated as 1 minus the joint probability; the lower the joint probability is, the higher the score is and therefore the more outlying the corresponding object is.

In Section 3.4, we mentioned the necessity of a prior in order to include all the possible cases in our model. We assume the same value for all the necessary priors. To choose the value of α, we calculate the number of instances one would hope to see if the data were uniformly distributed. Three parameters are considered for this estimation: the size of our data (5646), the number of bins in the discretization process (20), and the maximum number of parents a node can have on the BN. Given that the minimum number of parents is 0 and the maximum is 3, a reasonable number for α is 4. We also empirically tested with different values of α and found that the results are not sensitive to the choice of α.

5. DATA

5.1. MACHO Catalog

MACHO (Massive Compact Halo Object) is a survey that observed the sky, starting in 1992 July and ending in 1999, to detect microlensing events produced by Milky Way halo objects. Several tens of millions of stars were observed in the Large Magellanic Cloud (LMC), the Small Magellanic Cloud (SMC), and the Galactic bulge. The average number of observations per object is several hundred, with the center of the LMC being observed more frequently than the periphery. The reader can find a detailed description of MACHO in Alcock et al. (1997e).

Every light curve is described by 13 features corresponding to the blue nonstandard pass with a bandpass of 440–590 nm (see Pichara et al. 2012 for more details).

5.2. Training Set

The training set is composed of a subset of 5646 labeled observations from the MACHO catalog (Kim et al. 2011).[7] The constitution of the training set is presented in Table 1, and a representative example of each class light curve is shown in Figure 3.

The catalog comprises several sources from MACHO variable studies (Alcock et al. 1996, 1997a, 1997d, 1999), the MACHO microlensing studies (Alcock et al. 1997c, 1997e, 1997b; Thomas et al. 2005), and the LMC long-period variable study (Wood 2000). Quasars in the training set were collected from Blanco & Heathcote (1986); Schmidtke et al. (1999); Dobrzycki et al. (2002); Geha et al. (2003). Be stars were obtained from private communication with M. Geha (2009). The nonvariables were randomly chosen from the MACHO LMC database, and any previously known MACHO variables were removed from the nonvariable set.

6. RESULTS

In this section, we show how we applied the above methods to the MACHO catalog.

6.1. Performance Test

To prove the performance of our algorithm, we created a test set leaving one class out of the MACHO training set; we trained

[7] We collected these variables from the MACHO variable catalog found at http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=II/247.

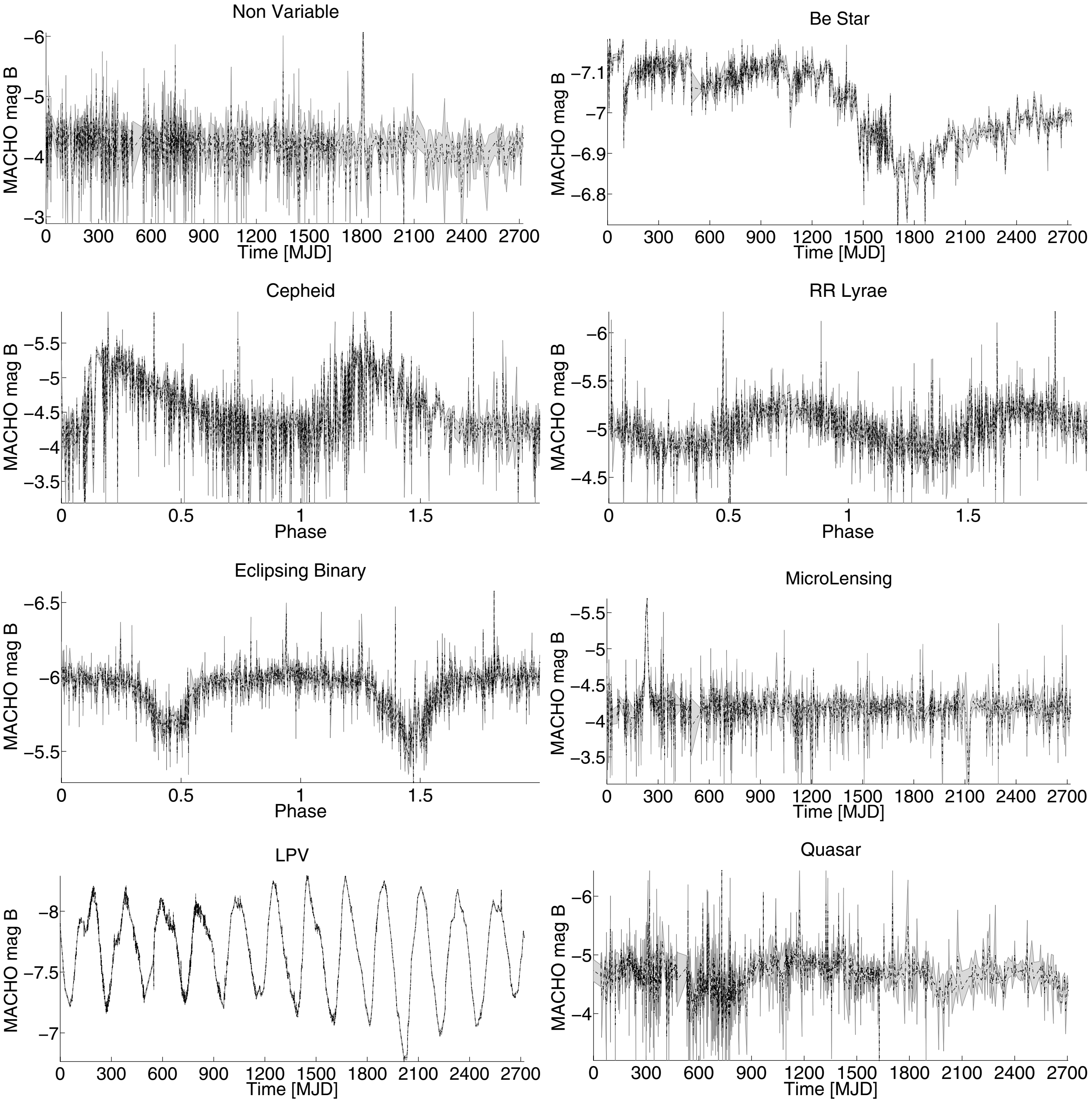
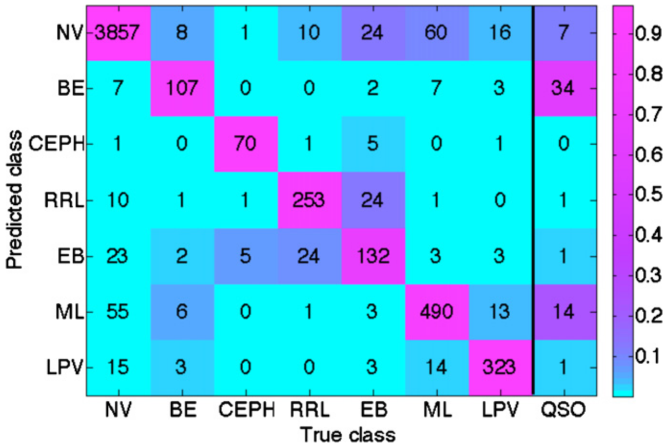


Figure 3. Example light curves of each class in the MACHO training set. The *x*-axis is the modified Julian date (MJD), and the *y*-axis is the MACHO *B* magnitude. Note that Cepheid, RR Lyrae, and eclipsing binary light curves are folded since they are periodic.

our algorithm with the remaining classes and considered the excluded class as unknown objects that we want to discover. In other words, we expected these light curves to have the highest outlierness score as they have never been seen by the model.

We performed three different tests, each time leaving out of the training set one of the classes: quasars, eclipsing binaries, and Be stars. The RF considered 500 trees, with $|F'| = \lfloor\sqrt{|F|}\rfloor$ features in every node.

Next, we present the results for the test leaving the quasars out of the training set. In order to visualize the voting database V, we present the average number of objects voted by the RF for each class in Figure 4. By using a color scale, we also show the average distribution of the votes among the different classes during the training phase and for the test class (quasars) during the testing phase (right vertical line). For example, when the RF is classifying an RR Lyrae, it has doubts mainly among nonvariables, eclipsing binaries, and the true class, RR Lyrae. This is shown in the colors along the vertical line labeled RRL. This hesitation is learned by the BN, and the relationship between classes is represented on a graph as shown in Figure 5. An RR Lyrae node is a child node of Cepheid and nonvariable nodes, meaning that when the light curve to be classified is from the RR Lyrae class, the voting vector will present high values in these other two classes. On the other hand, the Be star node is independent of the other classes, as expected.



Predicted class \ True class	NV	BE	CEPH	RRL	EB	ML	LPV	QSO
NV	3857	8	1	10	24	60	16	7
BE	7	107	0	0	2	7	3	34
CEPH	1	0	70	1	5	0	1	0
RRL	10	1	1	253	24	1	0	1
EB	23	2	5	24	132	3	3	1
ML	55	6	0	1	3	490	13	14
LPV	15	3	0	0	3	14	323	1

Figure 4. RF vote distribution (NV: nonvariable; BE: BE stars; CEPH: Cepheid; RRL: RR Lyrae; EB: eclipsing binaries; ML: microlensing, LPV: long-period variable; QSO: quasars).

(A color version of this figure is available in the online journal.)

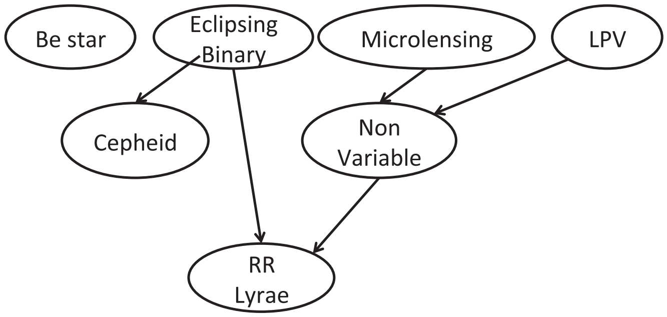


Figure 5. BN structure for the performance test.

After the algorithm training stage was completed, the outlier detection stage was performed. We first obtained from the trained RF a vector v_i for every object in the training set, quasars included. We then determined the joint probability and thus the outlier score of each v_i by using the already-learned BN. We were expecting the quasars to have the highest outlier scores and thus to find them on the top of the resulting outlier list. Figure 6 shows how objects of "known" classes present high joint probabilities while outlier objects (quasars) have the lowest values. Finally, the top left panel of Figure 7 represents our algorithm performance, comparing the imputed outlier (quasars) positions in the top outlier list with the ideal case result—the 58 quasars will be using the 58 first places in the outlier list. It can be seen that the top 40–60 outliers are quasars and all imputed outliers (quasars) are in the top 200 list. The same behavior is observed when we choose other classes as the outlier class, as shown in the top right and bottom panels of Figure 7.

6.2. Running on the Whole Data Set

Once we tested the accuracy of our method, we trained an RF (F-score $= 0.9080$) with the complete training set and learned a new BN. The same parameters of the performance test were used in this stage.

We ran our model on the whole MACHO data set (about 20 million light curves) to obtain a list of outlier candidates. Fortunately, the main computational cost of the algorithm occurs during the training phase, for which the model needs to build the RF and learn the BN structure and parameters. After training the model, performing the inference for a light curve takes a fraction of a second and can easily be repeated.

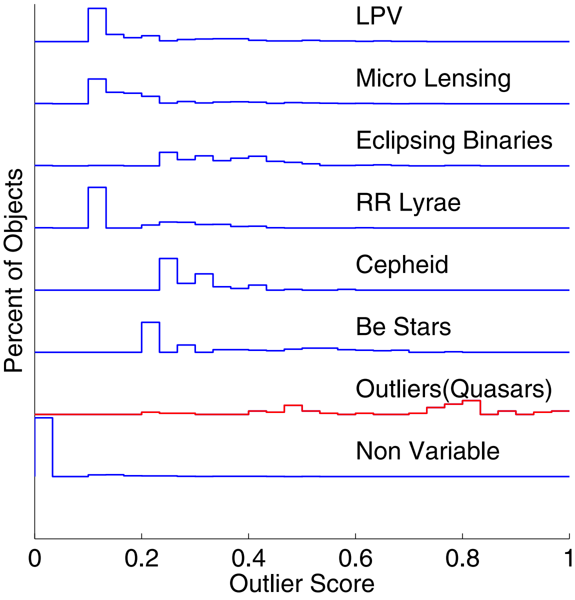


Figure 6. Stacked plot of the outlier score distribution for the performed test. Each layer represents the outlier score distribution of the objects of a class (blue lines show the training set classes and the red line the outlier class). The y-axis scale for each layer goes from 0 to 1, but it was removed for visual clarity.

(A color version of this figure is available in the online journal.)

6.2.1. Removal of Spurious Outliers

Figure 8 shows some of the outliers we obtained from this first iteration. The top left and right outliers in Figure 8 are characterized by having one-day periods, while the bottom right has a period of approximately a year. This is probably caused by MACHO's nightly and seasonal observational pattern and not by an intrinsic anomalous behavior. We also faced other kinds of artifacts, like the outlier in the bottom left panel of Figure 8, which is obviously due to some instrumentation problems—this behavior at the beginning of the light curve appeared in many light curves.

In order to remove the spurious outliers, we perform the following steps.

1. Filter all outlier candidates that have periods very close to a sidereal day or a year. There is no doubt that those light curves exhibit strange behavior owing to variable seeing conditions during the night or seasonal aliases.
2. We run the entire analysis in the MACHO red nonstandard bandpass. MACHO was observed in two bandpasses simultaneously, and therefore there are corresponding red-band light curves for each object. For every outlier candidate that is not in the top 20,000 list of the equivalent list in the red candidate list, we consider it as an artifact/spurious and therefore remove it from the candidate list.
3. We visually inspect all candidates and group those that are obviously spurious, like the examples in Figure 8, into groups of similar shapes and behaviors. We add these new classes to the training set, retrain, and then predict outliers again as explained above.
4. Repeat previous steps until finding no artifacts on the top outlier list.

We expect that once we filter the artifacts, the true outliers will be the only ones remaining.

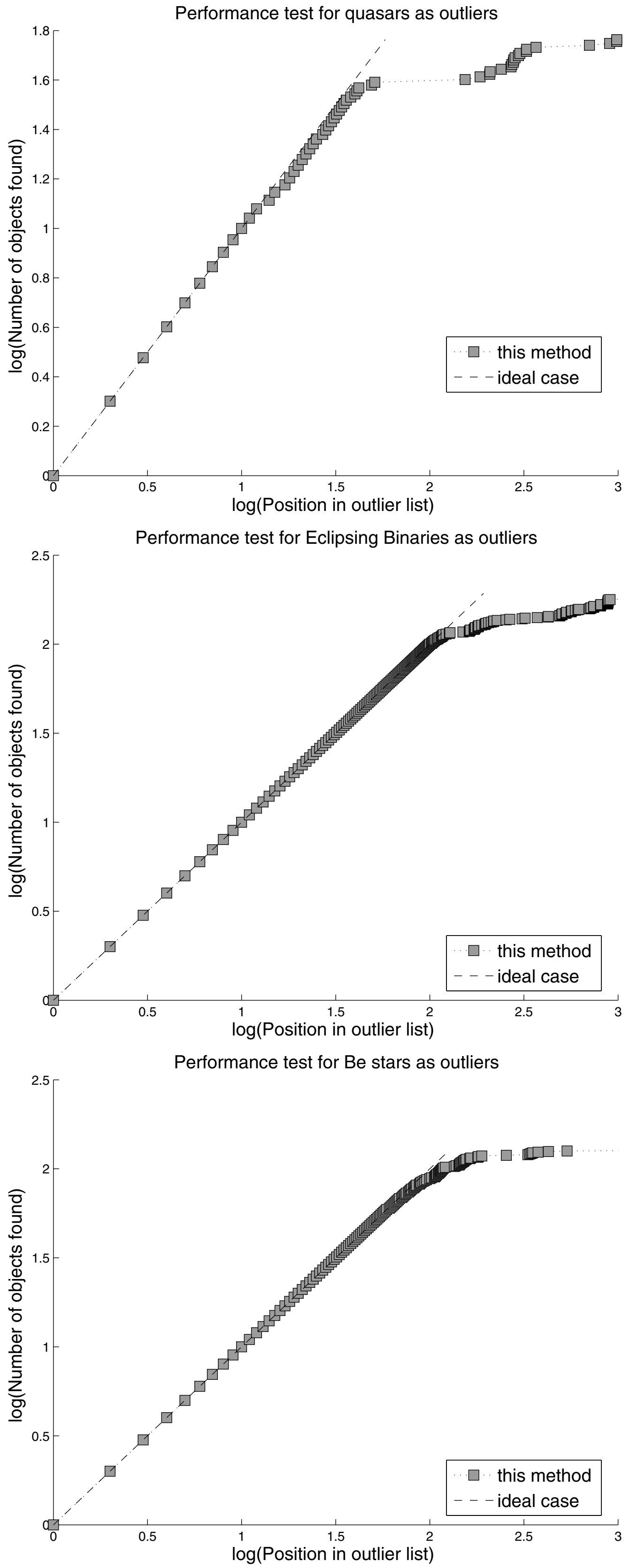


Figure 7. Performance test results for quasars, eclipsing binaries, and Be stars as outliers. The dashed line represents the ideal result, where the class left out uses the top positions in the outlier list. Gray squares show the actual positions obtained.

7. POST ANALYSIS

As a first step, we visually inspected all the candidates starting from the top of the list ("strongest" outliers) and moving our way to the "weakest" outliers. We determined that about 4000 candidates was a good number of candidates to start, since candidates beyond this point either were not showing any significant variation or had low signal-to-noise ratio (S/N) and therefore were not interesting.

As a second step, we cross-matched our candidates with other astronomical catalogs of known types, or catalogs with additional contextual information. Some of these catalogs are collections of known types; for example, LMC Long Period Variables (Fraser et al. 2008) is a collection of long-period variables from LMC. On the other hand, catalogs like *XMM-Newton* (Watson et al. 2009) contain X-ray information, which can be useful to further understand the nature of the candidates. Having additional information for some of the outlier candidates could be helpful to identify the nature of these objects. Table 3 summarizes all the catalogs used in the analysis and the resulting cross-matched numbers ($N_{\text{x-matched}}$).

The fact that some of the outlier candidates appear in catalogs of known objects, as is shown in Table 3, could be explained by the following reasons.

1. Known classes with a small number of objects were not included in our original training set (i.e., cataclysmic variables, R Coronae Borealis, etc.). Since these are rare classes, we were expecting to find more objects of their kind.
2. The objects in these catalogs were mislabeled or incorrectly classified. Many of these catalogs are guided by algorithms or done automatically, so unavoidably they contain errors. Even when humans are involved in the classification, biases are always present. These "errors" should present themselves as outliers in our final analysis. Indeed, 45 of our outliers that were labeled as eclipsing binaries, Cepheids, or RR Lyrae in other catalogs do not have the characteristics or the light curve shapes of these classes and therefore were flagged as outliers.
3. The features considered in this work and the features used by the other catalogs are not consistent. For example, the period of `MACHO_77.7428.190` is 906.3559 days, while in Soszynski et al. (2008) it is 0.2843359 days. Because of this, this light curve does not appear to be an RRL in our model, and therefore it is identified as an outlier. It is known that uncertainties in features could result in low-quality classification and consequently erroneous outlier predictions. Dealing with feature uncertainties is a topic of future work.
4. The S/N of the light curves is survey-dependent, and therefore features that depend on the actual amplitude of the variability vary from catalog to catalog. For example, if a catalog is compiled using a survey that is more sensitive than ours, the fainter objects are indistinguishable from the nonvariables in our database even if it is a true known variable. Moreover, as described above, low-S/N light curves have uncertain features and therefore higher probability of being false positive.

Most of these reasons can be attributed to the lack of a perfect training set and high-quality features. Because our method is based on a supervised classification, the results depend heavily on the choice of these representative objects. In an ideal scenario, one would compile a training set that contains every possible

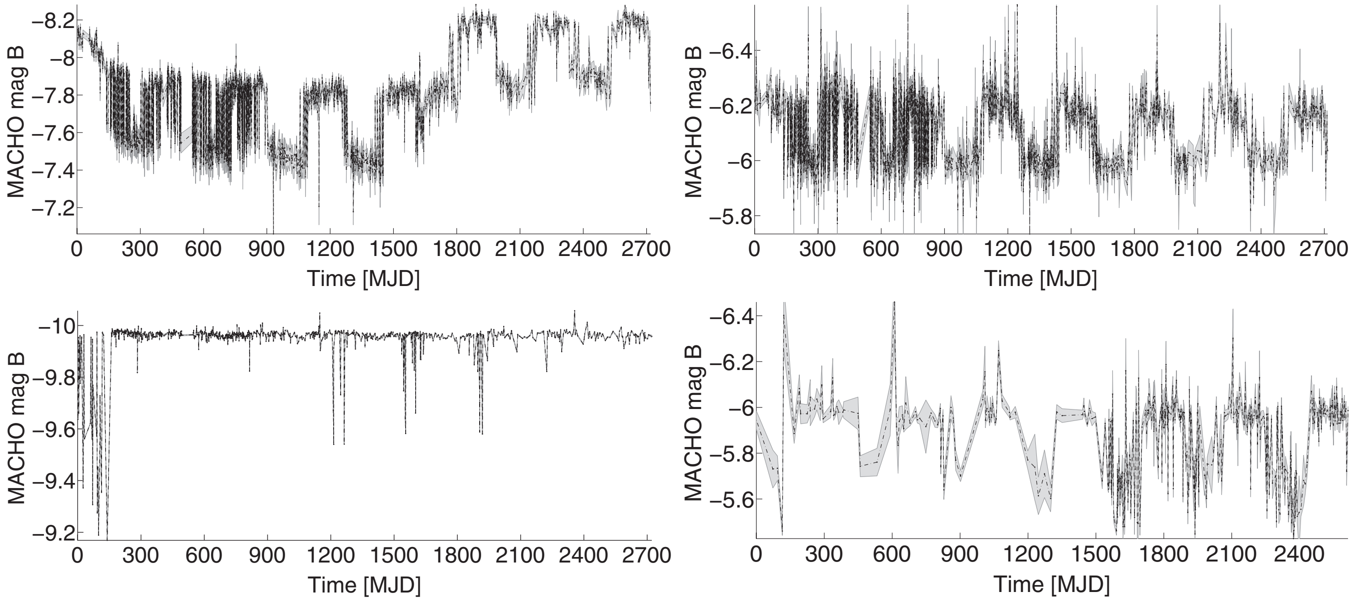


Figure 8. Top left panel: one-day period artifact MACHO_77.7187.271; top right panel: one-day period artifact MACHO_79.4780.358; bottom left panel: sampling artifact MACHO_5.5010.986; bottom right panel: 370 day period MACHO_49.5899.715.

Table 3
Catalogs Used for Post-analysis

Catalog	Reference	Number of objects in catalog	$N_{\mathrm{x-matched}}$
LMC LPVs from MACHO	Fraser et al. (2008)	56,453	52[a]
XMM-Newton	Watson et al. (2009)	262,902	13
ROSAT All-Sky Bright Source Catalogue (1RXS)	Voges et al. (1999)	18,806	2
LMC Blue variable stars from MACHO	Keller et al. (2002)	1280	91
OGLE eclipsing binaries in LMC	Wyrzykowski et al. (2003)	2720	29
OGLE RR Lyrae in LMC	Soszynski et al. (2003)	7661	8
LMC Cepheids in OGLE and MACHO data	Poleski (2008)	2946	8
OGLE+2MASS+DENIS LPV in Magellanic Clouds	Groenewegen (2004)	2919	9
Variable Stars in the Large Magellanic Clouds	Alcock et al. (2004)	21474	334
Machine-learned ASAS Classification Cat. (MACC)	Richards et al. (2012)	50124	5
QSO Candidates in the MACHO LMC database	Kim et al. (2012)	2566	51
EROS Periodic Variable Candidates	Kim et al. (2014)	150,115	432
Type II and anomalous Cepheids in LMC	Soszynski et al. (2008)	286	19
OGLE Variables in Magellanic Clouds	Ita et al. (2004)	8852	134
GCVS, Vol. V.: Extragalactic Variable Stars	Artyukhina et al. (1996)	10979	74
High proper-motion stars from MACHO astrometry	Alcock et al. (2001)	154	0

Note. [a] 52 were types 0, 9, or no types in this paper.

Table 4
The Others: Examples of New Variability Classes and Individual Outliers

Class	MACHO id	R.A.	Decl.	Period [days]	V	R	Color	S/N
Class A	82.8887.471	5.59031	−69.2956	657.19	19.78	19.28	0.49	1.53
Class A	82.9009.834	5.59633	−69.2722	525.75	20.25	19.71	0.536	2.38
Class A	82.9009.1850	5.59655	−69.2762	525.75	21.04	20.96	0.08	1.66
Class A	82.8887.2395	5.59106	−69.2954	876.25	21.04	21.20	−0.16	1.59
Class B	56.5178.29	5.19911	−66.5471	363.00	16.50	17.02	−0.51	4.44
Class B	44.1616.257	4.84559	−70.0673	871.32	16.23	16.704	−0.47	3.84
Class B	35.7272.13	5.42992	−72.127	374.98	16.23	16.70	−0.47	5.63
Class B	48.2864.67	4.96026	−67.5326	872.94	17.00	17.53	−0.52	2.98
Class B	82.8284.126	5.51805	−69.202	438.12	17.56	17.61	−0.04	2.34
Class C	17.2711.26	4.9556	−69.6723	680.70	15.41	15.87	−0.46	9.14
Class C	82.8283.41	5.5218	−69.2594	525.75	15.07	15.66	−0.59	8.53
Class C	62.7361.30	5.4249	−66.2181	848.30	16.38	16.65	−0.27	5.44
Individual Outlier	13.5835.11	5.2742	−71.0974	296.98	14.85	15.21	−0.36	51.52
Individual Outlier	18.2478.9	4.9342	−69.0323	226.90	14.76	15.23	−0.46	36.69
Individual Outlier	78.6462.561	5.3366	−69.6743	678.95	18.11	17.91	0.20	7.02
Individual Outlier	62.7241.19	5.4114	−66.1581	636.23	16.16	16.34	−0.18	52.51

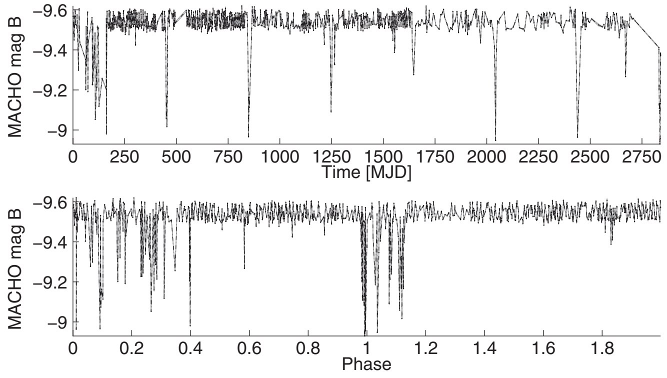


Figure 9. Top panel: eclipsing Cepheid MACHO_6.6454.5; bottom panel: its folded light curve.

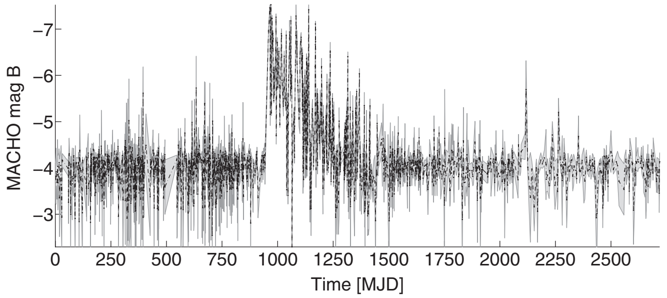


Figure 10. Nova-like variable MACHO_77.7546.2744.

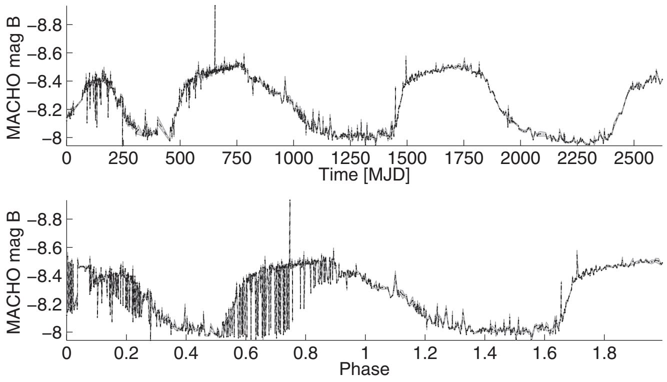


Figure 11. Blue variable MACHO_81.9727.662.

known object with high-quality features. In our case, we started with a trustworthy training set that was missing some of the known but rare types. This served as a blind test, since some of these types were never presented to the method, never trained with them, and therefore should have been discovered by our method. As expected, we recovered most of these objects in the candidate list.

As a third step, we examined the color–magnitude diagram (CMD) of the candidate list and identified regions where objects were most likely from a known type. One of the advantages of the LMC is that all stellar populations are at essentially the same distance, and thus we can use CMDs as an additional way to separate and identify the sources. Figure 14 shows the CMD for the outliers.

As a fourth step, we grouped the outliers into sets based on the morphology of the light curves. Here we present the most interesting subgroups, some of which are known but rare classes, while others do not obviously belong to any known class of objects.

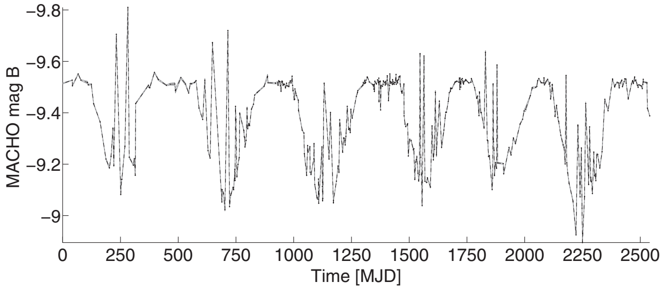


Figure 12. X-ray binary MACHO_61.9045.32.

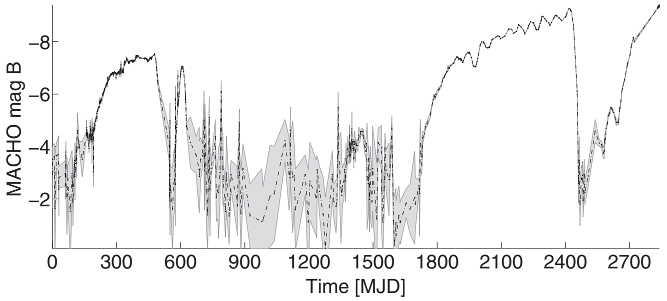


Figure 13. R Coronae Borealis MACHO_6.6696.60.

1. *Eclipsing Cepheid:* Eclipsing Cepheids have been discussed in papers of the MACHO, OGLE, and EROS-2 surveys (Alcock et al. 2002; Marconi et al. 2013; Cassisi & Salaris 2011). These objects are Cepheids in binary systems where there are flux drops during the pulsating cycle caused by the transit of a companion star. Although it is known that 50% of Galactic Cepheids are in binary systems, only about 20 such Cepheids are known in the LMC, which is mainly due to their faint magnitudes caused by the distance to the LMC. Recently, Pietrzyński et al. (2010) have used such a system to limit the distance uncertainty to the LMC, so finding such systems is very valuable for precision cosmology. By simply looking through our catalog of outliers, we found few objects of this kind. Figure 9 shows one of these examples.
2. *Cataclysmic Variables (CVs):* Another interesting group of outliers are CVs, novae, or novae-like-looking objects. Because there are no unified variability characteristics, this group was not included in the training set, and therefore there are few CVs in our candidate list. These objects can increase more than 20 mag, becoming approximately 10^8 times brighter. Novae and recurrent novae are close binary systems that are variable owing to explosions on their surfaces. The eruptions can last from a few days to almost a year and can be quasi-periodic as the recurrent novae (Schaefer 2010; Knigge 2011). This is a subject of extensive research, and recently the interests focused on superluminous supernovae (Quimby et al. 2011). Figure 10 shows `MACHO_77.7546.2744`, one example of this class, where the change in magnitude is 2.5 and the relaxation time is about a year. Our candidate list contains a few dozen of these objects; nevertheless, some of them are already known, such as those presented in Shafter (2013).
3. *Blue Variables:* the class coined blue variables is a generic class without a unified light curve morphology or features. Because of this, we did not include such a class in the training set. Keller & Wood (2002) proposed that the

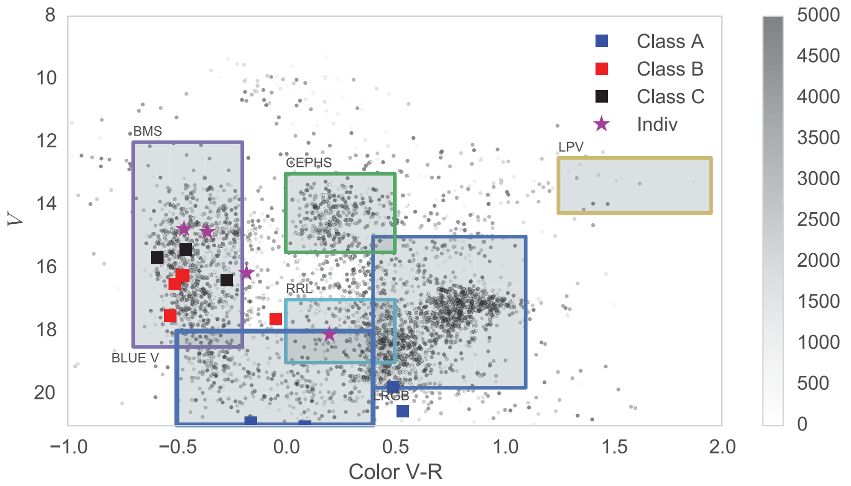


Figure 14. CMD of all the outliers. The outlier rank is indicated by the color of each data point. The bluer, the higher the outlier score. Black boxes mark the location of blue main sequence (BMS), lower red giant branch (LRGB), long-period variables (LPV), RR Lyrae (LLR), and Cepheid (CEPH).

(A color version of this figure is available in the online journal.)

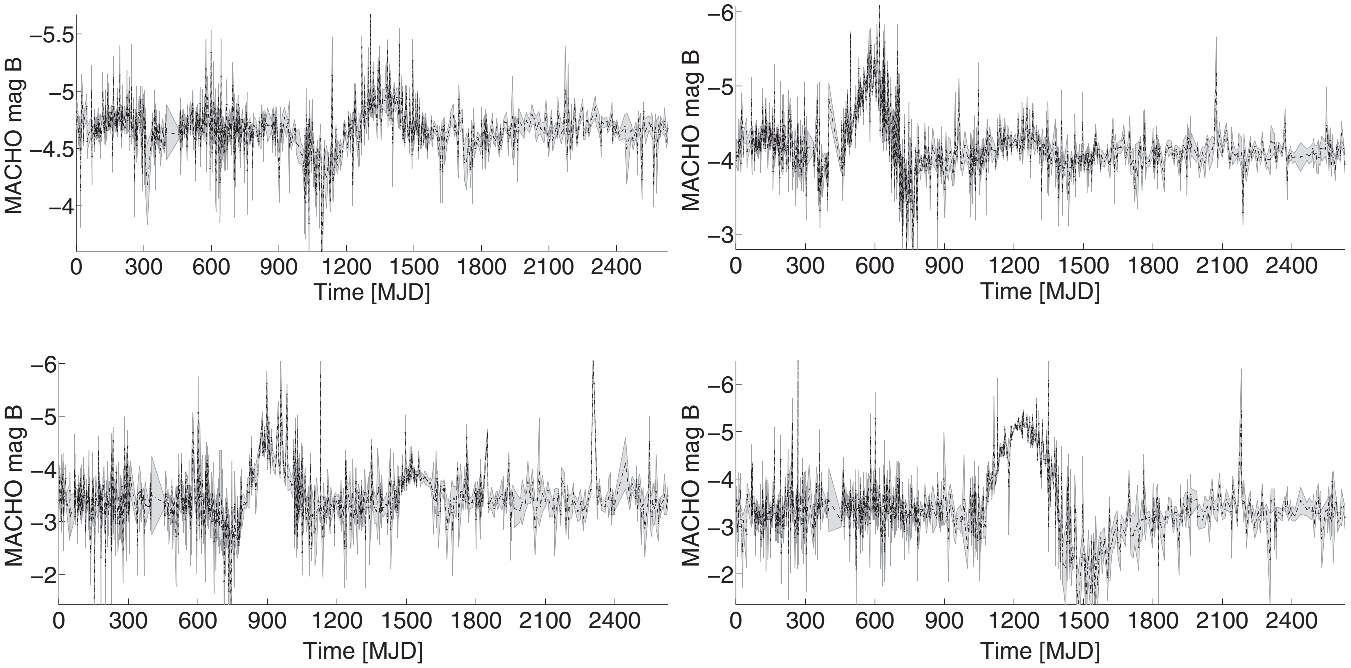


Figure 15. Top left panel: Class_A MACHO_82.8887.471; top right panel: Class_A MACHO_82.9009.834; bottom left panel: Class_A MACHO_82.9009.1850; bottom right panel: Class_A MACHO_82.8887.2395.

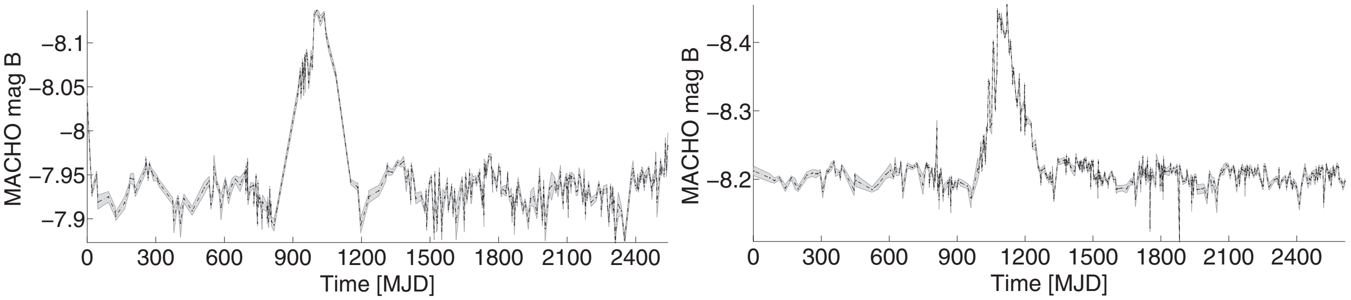


Figure 16. Left panel: Class_B MACHO_56.5178.29; right panel: Class_B MACHO_44.1616.257.

variability of these stars is the result of processes related to the establishment, maintenance, and dissipation of the Be disk. The emission that characterizes Be stars originates in a gaseous circumstellar quasi-Keplerian disk. These objects appear to be blue and are simply variable. Sixty-eight of our candidates fall into this category. An example of such a light curve is shown in Figure 11, and the locations of all the members in the CMD are shown in Figure 14.

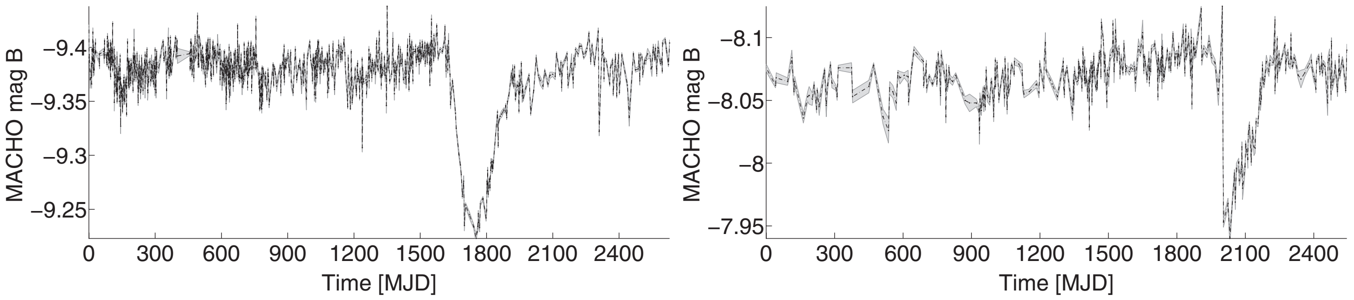


Figure 17. Left panel: Class_C MACHO_82.8283.41; right panel: Class_C MACHO_62.7361.30.

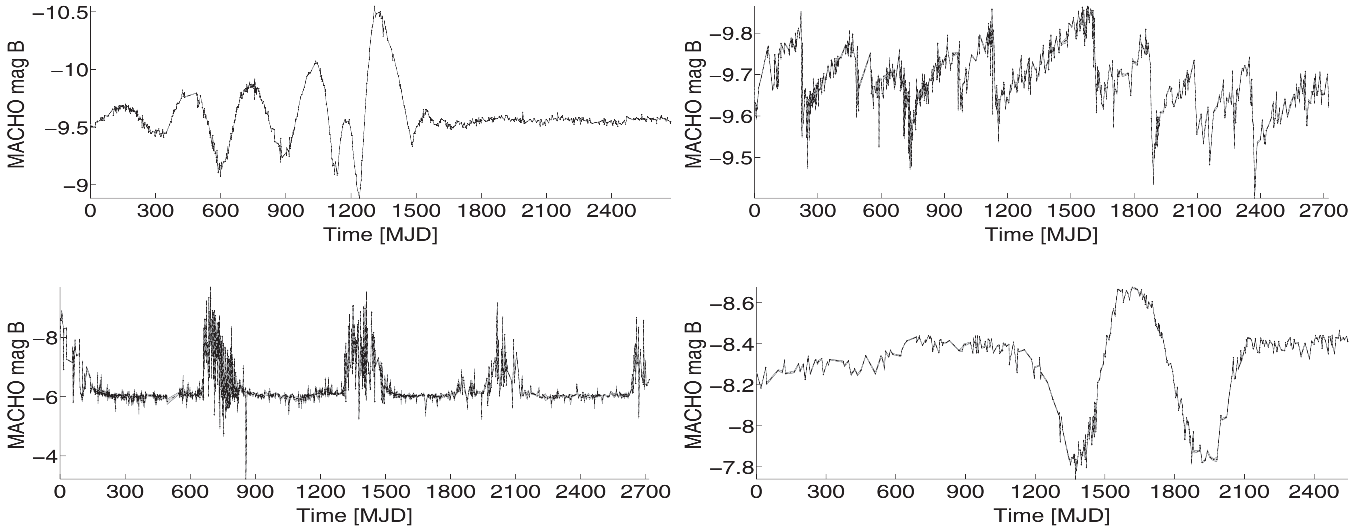


Figure 18. Top left panel: outlier MACHO_13.5835.11; top right panel: outlier MACHO_18.2478.9; bottom left panel: outlier MACHO_78.6462.561; bottom right panel: outlier MACHO_62.7241.19.

4. *X-ray Sources:* there are 2 sources cross-matched with the *ROSAT* all-sky survey bright source catalog (Voges et al. 1999) and 13 with the second *XMM-Newton* serendipitous source catalog (Watson et al. 2009). Among these X-ray sources, `MACHO_61.9045.32` is a confirmed high-mass X-ray binary (Liu et al. 2005) hosting a radio pulsar (Ridley et al. 2013), but the other 14 counterparts are not carefully studied for their X-ray origins. These remaining objects are interesting sources since they show strong optical variability, either periodic or nonperiodic, and X-ray emission simultaneously. They could be either W UMa-type contact binaries, X-ray binaries, or other types of X-ray emitters (e.g., see Ness et al. 2002; Chen et al. 2006; Liu et al. 2007 and references therein). Particularly, X-ray binaries are most interesting sources since they are known to host either neutron stars or black holes (i.e., accretor) together with a companion star. Their X-ray emission is caused by accreting material falling from the companion star into the accretor (van den Heuvel et al. 1992; Done et al. 2007). Thus, studying X-ray binaries helps us to understand the process of accretion and the fundamental physics of the binaries such as mass, radius, orbit, and jets (e.g., see van der Klis 2000; Fender et al. 2004). Figure 12 shows one representative example.
5. *R Coronae Borealis:* within our outliers we identified one object belonging to one of the most rare and interesting classes among the variable stars. `MACHO_6.6696.60` is an R Coronae Borealis star. These kinds of objects are yellow supergiant stars whose atmospheres are carbon-rich and extremely hydrogen-deficient. This causes irregular intervals of dust-formation episodes that result in a drop in brightness of up to 8 mag in a short period (Clayton 1996). An example of this type of light curve is shown in Figure 13.
6. *The Others:* undoubtedly, there are many variable classes, and it is out of the scope of this work to analyze and comment on every outlier from our list. Our goal was to find novel objects that have not been identified before. For this end, we first ran a clustering algorithm on all the candidates, then visually inspected all the light curves that are not in the categories mentioned above, and identified a few classes of objects and a few individual objects that could not be assigned to known classes. We show three classes and four individual outliers in Table 4, in Figures 15, 16, 17, 18, and also in the CMD in Figure 14.

   Nevertheless, we had to perform a more specific analysis for outliers in Class A. We noticed that the objects belonging to this class are neighbors (they are located in the same field, number 82), and therefore it is likely that the perturbation on the light curves was caused by a high-proper-motion star moving close to these sources. In order to confirm or reject this hypothesis, we calculated the distance between these objects and the time differences of the peaks of the variation. The time difference of the variation was on average 400 days, but the objects were $\sim 100''$apart. Since typical proper motions are less than a few arsec per year, the hypothesis was rejected. Objects of Class A are consequently good candidates to conform a new variability class.

8. CONCLUSIONS

The generation of precise, large, and complete sky surveys in recent years has increased the need of developing automated analysis tools to process this tremendous amount of data. These tools should help astronomers to classify stars, characterize objects, and detect anomalies, among other applications. In this paper, we presented an algorithm based on a supervised classifier mechanism that enables us to discover outliers in catalogs of light curves. To do so, we trained an RF classifier and used a BN to obtain the joint probability distribution, which was used for our outlierness score. Different from existing methods, our work comprises a supervised algorithm where all the available information is used to our advantage.

Since the amount of data to be processed is huge, one could have expected a high computational complexity and the overtaking of the resources. Nevertheless, our algorithm is only expensive in the training stage and is extremely fast in the unknown light curve analysis, allowing us to explore very large data sets. Furthermore, our method is not only restricted to astronomical problems and could be applied to any database where anomaly detection is necessary.

The results from the application of our work on catalogs of classified periodic stars from the MACHO project are encouraging and establish that our method correctly identifies light curves that do not belong to these catalogs as outliers.

We have identified light curves that were artifacts because of instrumental, mechanical, electronic, or human errors and about 4000 light curves that emerged as intrinsic. After cross-matching these candidates with the available catalogs, we found known but rare objects among our outliers and also objects that did not have previous information. By performing a clustering, we classified some of them as new variability classes and others as intriguing unique outliers. As future work these objects will be followed up using spectroscopy in order to characterize them and identify them with new observations. We hope that by doing this analysis we would be able to find more of these objects and turn our isolated outliers into new known variability classes.

On the other hand, we are planning to improve our algorithm in the future by creating new robust features and by constructing a more complete and large training set. Furthermore, we aim to apply our algorithm to different large sky surveys such as EROS (Ansari 2004), Pan-Starrs (Hodapp et al. 2004), and, when finished, LSST (Tyson et al. 2002).

Finally, in order to help astronomers, we are planning a full release of software that will include feature calculation of the light curves and the application of our algorithm as a downloadable software and as an online tool and web services in the near future.

We thank M. Stockle and D. Acuña for helpful comments and advice. The analysis in this paper has been done using the Odyssey cluster offered by the FAS Research Computing Group at Harvard. This work is supported by Vicerrectoría de Investigación (VRI) from Pontificia Universidad Católica de Chile and by the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC 12009, awarded to The Millennium Institute of Astrophysics.

REFERENCES

Agarwal, D. 2005, in Proc. 5th IEEE Int. Conf. Data Mining IEEE Computer Society, ed. D. Xin, J. Han, X. Li et al. (Los Alamitos, CA: IEEE Computer Society), 26

Aggarwal, C. C., & Yu, P. S. 2001, ACM SIGMOD Record, 30, 37

Alcock, C., Allsman, R. A., Alves, D., et al. 1997a, ApJ, 482, 89

Alcock, C., Allsman, R. A., Alves, D., et al. 1997b, ApJ, 486, 697

Alcock, C., Allsman, R. A., Alves, D., et al. 1997c, ApJL, 491, L11

Alcock, C., Allsman, R. A., Alves, D., et al. 1997d, AJ, 114, 326

Alcock, C., Allsman, R. A., Alves, D., et al. 1997e, ApJ, 479, 119

Alcock, C., Allsman, R. A., Alves, D. R., et al. 1999, AJ, 117, 920

Alcock, C., Allsman, R. A., Alves, D. R., et al. 2001, ApJ, 562, 337

Alcock, C., Allsman, R. A., Alves, D. R., et al. 2002, ApJ, 573, 338

Alcock, C., Allsman, R. A., Axelrod, T. S., et al. 1996, AJ, 111, 1146

Alcock, C., Alves, D. R., Axelrod, T. S., et al. 2004, AJ, 127, 334

Ansari, R. 2004, in International Conference on Cosmic Rays and Dark Matter, ed. Y. Muraki (Tokyo: Universal Academy Press), 1

Arning, A., Agrawal, R., & Raghavan, P. 1996, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ed. E. Simoudis, J. Han, & U. M. Fayyad (Palo Alto, CA: AAAI Press), 164

Artyukhina, N., Durlevich, O., Frolov, M., et al. 1996, VizieR Online Data Catalog, 2205, 0

Bernard, S., Heutte, L., & Adam, S. 2008, in Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, ed. D.-S. Huang, D. C. Wunsch II, D. S. Levine, & K.-H. Jo (Berlin: Springer), 430

Bhattacharyya, S., Richards, J. W., Rice, J., et al. 2012, in Statistical Challenges in Modern Astronomy V, Vol. 902 (Berlin: Springer), 483

Bishop, C. 1994, in Proc. IEEE Conf. Vision, Image and Signal Processing (London: Springer), 217

Blanco, V. M., & Heathcote, S. 1986, PASP, 98, 635

Breiman, L. 2001, Mach. Learn., 45, 5

Breunig, M., Kriegel, H., Ng, R., & Sander, J. 2000, in Proc. 2000 ACM SIGMOD Int. Conf. Management of Data, ed. J. M. Zytkov & J. Rauch (London: Springer), 93

Cassisi, S., & Salaris, M. 2011, ApJL, 728, L43

Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Comput. Surveys, 41, 1

Chen, W. P., Sanchawala, K., & Chiu, M. C. 2006, AJ, 131, 990

Clayton, G. C. 1996, PASP, 225

Cooper, G., & Herskovits, E. 1992, Mach. Learn., 9, 309

Dobrzycki, A., Groot, P. J., Macri, L. M., & Stanek, K. Z. 2002, ApJL, 569, L15

Done, C., Gierliński, M., & Kubota, A. 2007, A&ARv, 15, 1

Eskin, E. 2000, in Proc. 7th Int. Conf. on Machine Learning, ed. P. Langley (San Francisco, CA: Morgan Kaufmann Publishers, Inc.), 255

Fender, R. P., Belloni, T. M., & Gallo, E. 2004, MNRAS, 355, 1105

Fraser, O. J., Hawley, S. L., & Cook, K. H. 2008, AJ, 136, 1242

Geha, M., Alcock, C., Allsman, R. A., et al. 2003, AJ, 125, 1

Geurts, P., Ernst, D., & Wehenkel, L. 2006, Mach. Learn., 63, 3

Gibbons, P. B., & Matias, Y. 1998, in SIGMOD Rec., 27, 331

Groenewegen, M. 2004, A&A, 425, 595

Grubb, Frank, F. E. 1969, Technometrics, 11, 1

He, J., & Carbonell, J. 2006, in Proceedings of the 10th International Symposium on Artificial Intelligence and Mathematics, ed. M. C. Golumbic (London: Springer)

Henrion, M., Hand, D. J., Gandy, A., & Mortlock, D. J. 2013, Stat. Anal. Data Mining, 6, 53

Herschel, J. 1857, Outlines of Astronomy (Philadelphia, PA: Blanchard and Lea), 268

Hodapp, K. W., Kaiser, N., Aussel, H., et al. 2004, AN, 325, 636

Hodge, V., & Austin, J. 2004, Artif. Intell. Rev., 22, 85

Ita, Y., Tanabé, T., Matsunaga, N., et al. 2004, MNRAS, 353, 705

Jin, W., Tung, A., & Han, J. 2001, in Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, ed. T. Fawcett (New York: ACM), 293

John, G. 1995, in Proc. 1st Int. Conf. Knowledge Discovery and Data Mining, ed. U. M. Fayyad & R. Uthurusamy (Palo Alto, CA: AAAI Press), 174

Keller, S. C., Bessell, M. S., Cook, K. H., Geha, M., & Syphers, D. 2002, AJ, 124, 2039

Keller, S. C., Schmidt, B. P., Bessell, M. S., et al. 2007, PASA, 24, 1

Keller, S. C., & Wood, P. R. 2002, ApJ, 578, 144

Kim, D.-W., Protopapas, P., Bailer-Jones, C., et al. 2014, A&A, 566, A43

Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68

Kim, D.-W., Protopapas, P., Trichas, M., et al. 2012, ApJ, 747, 107

Knigge, C. 2011, in ASP Conf. Ser. 447, Evolution of Compact Binaries, ed. L. Schmidtobreick, M. R. Schreiber, & C. Tappert (San Francisco, CA: ASP), 3

Knorr, E., & Ng, R. 1998, in Proc. VLDB Conf., New York, USA, ed. A. Gupta, O. Shmueli, & Widom (San Francisco, CA: Morgan Kaufmann Publishers, Inc.), 392

Koller, D., & Friedman, N. I. R. 2009, Probabilistic Graphical Models (Cambridge, MA: MIT Press)

Kou, Y., Lu, C., Sirwongwattana, S., & Huang, Y. 2004, in Proc. IEEE Int. Conf. Networking, Sensing and Control (Los Alamitos, CA: IEEE Computer Society), 749

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2005, A&A, 442, 1135

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2007, A&A, 469, 807

Marconi, M., Molinaro, R., Bono, G., et al. 2013, ApJL, 768, L6

Monti, S., & Cooper, G. F. 1998, in Proc. 14th Conf. Uncertainty in Artificial Intelligence, ed. G. Cooper & S. Moral (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 404

Nairac, A., Towsend, N., Carr, R., et al. 1999, Intgr. Comput. Aided Eng., 6, 53

Ness, J.-U., Schmitt, J. H. M. M., Burwitz, V., Mewe, R., & Predehl, P. 2002, A&A, 387, 1032

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. 2003, in Proc. 19th Int. Conf. Data Engineering, ed. U. Dayal, K. Ramaritham, & T. M. Vijayaraman (Los Alamitos, CA: IEEE Computer Society), 315

Penzias, A., & Wilson, R. 1965, ApJ, 142, 419

Pichara, K., & Protopapas, P. 2013, ApJ, 777, 83

Pichara, K., Protopapas, P., Kim, D.-W., Marquette, J.-B., & Tisserand, P. 2012, MNRAS, 427, 1284

Pichara, K., & Soto, A. 2011, Intell. Data Anal., 15, 151

Pichara, K., Soto, A., & Araneda, A. 2008, in Advances in Artificial Intelligence-IBERAMIA, ed. H. Geffner, R. Prada, I. Machado, & D. Nuno (London: Springer), 163

Pietrzyński, G., Thompson, I. B., Gieren, W., et al. 2010, Natur, 468, 542

Poleski, R. 2008, AcA, 58, 313

Protopapas, P., Giammarco, J., Faccioli, L., et al. 2006, MNRAS, 369, 677

Quimby, R. M., Kulkarni, S. R., Kasliwal, M. M., et al. 2011, Natur, 474, 487

Ramaswamy, S., Rastogi, R., & Shim, K. 2000, in Proc. ACM SIGMOD Conf. Management of Data, Dallas, TX, ed. W. Chen, J. F. Naughton, & P. A. Bernstein (New York: ACM), 427

Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. 2008, Mach. Learn., 74, 281

Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32

Ridley, J. P., Crawford, F., Lorimer, D. R., et al. 2013, MNRAS, 433, 138

Schaefer, B. E. 2010, ApJS, 187, 275

Schmidtke, P. C., Cowley, A. P., Crane, J. D., et al. 1999, AJ, 117, 927

Seidl, T., Müller, E., Assent, I., & Steinhausen, U. 2009, in Uncertainty Management in Information Systems, ed. C. Koch, B. Konig-Reis, V. Markel, & M. van Keulen (Dagstuhl, Germany: Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik)

Serio, G. F., Manara, A., Sicoli, P., & Bottke, W. F. 2002, in Giuseppe Piazzi and the discovery of Ceres (Tucson, AZ: Univ. Arizona Press)

Shafter, A. W. 2013, AJ, 145, 117

Son, C., Cho, S., & Yoo, J. 2009, in Management Enabling the Future Internet for Changing Business and New Computing Services, Vol. 5787, ed. C. S. Hong, T. Tonouchi, Y. Ma, & C.-S. Chao (Berlin: Springer), 291

Soszynski, I., Udalski, A., Kubiak, M., Zebrun, K., & Szewczyk, O. 2003, AcA, 53, 93

Soszynski, I., Udalski, A., Szymanski, M., et al. 2008, AcA, 58, 293

Thomas, C. L., Griest, K., Popowski, P., et al. 2005, ApJ, 631, 906

Tyson, J. A., Collaboration, L., Labs, B., Technologies, L., & Hill, M. 2002, Astron. Teles. Instrum., 4836, 10

van den Heuvel, E. P. J., Bhattacharya, D., Nomoto, K., & Rappaport, S. A. 1992, A&A, 262, 97

van der Klis, M. 2000, A&A, 38, 717

Voges, W., Aschenbach, B., Boller, T., et al. 1999, A&A, 349, 389

Watson, M., Schröder, A., Fyfe, D., et al. 2009, A&A, 493, 339

Wood, P. R. 2000, PASP, 17, 18

Wyrzykowski, L., Udalski, A., Kubiak, M., et al. 2003, AcA, 53, 1

Xiong, L., Poczos, B., Connolly, A., & Schneider, J. 2010, Anomaly Detection for Astronomical Data

Yang, H., Xie, F., & Lu, Y. 2006, in Fuzzy Systems and Knowledge Discovery, Vol. 4223, ed. L. Wang, L. Juao, G. Shi, X. Li, & J. Liu (Berlin: Springer), 1082

Zhang, T., Ramakrishnan, R., & Livny, M. 1996, in Proc. ACM SIGMOD Int. Conf. Management of Data, ed. J. M. Zytkov & J. Rauch (London: Springer), 103