PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

SCHOOL OF ENGINEERING

SUPERVISED DETECTION OF ANOMALOUS LIGHT-CURVES IN MASSIVE ASTRONOMICAL CATALOGS

ISADORA TATIANA NUN BITRÁN

Thesis submitted to the Office of Research and Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science in Engineering

Advisor:

KARIM PICHARA BAKSAI

Santiago de Chile, April 2014

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

SCHOOL OF ENGINEERING

SUPERVISED DETECTION OF ANOMALOUS LIGHT-CURVES IN MASSIVE ASTRONOMICAL CATALOGS

ISADORA TATIANA NUN BITRÁN

Members of the Committee:
KARIM PICHARA BAKSAI
LORETO VALENZUELA ROEDIGER
.....

ROLANDO DUNNER PLANELLA
LEONARDO VANZI

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, April 2014

To my family, friends and Ruby

ACKNOWLEDGEMENTS

I would like to thank my advisor Karim Pichara, for strongly encouraging me to pursue my Master studies. Also, for trusting me from the very beginning and for his guidance through all this process.

I would also like to thank Pavlos Protopapas for proving to me that is possible to be both extraordinary as a scientist and a human being. For his invaluable help and for his tremendous contribution and commitment to this work.

Special thanks to Daniel Acuña and Marcelo Stöckle for sharing with me their computational knowledge and keeping me company during moments of despair.

In addition, I wish to express my gratitude to the Institute for Applied Computational Science at Harvard for giving me the opportunity to take part on their research program.

Finally, an honorable mention goes to my whole family, specially to my parents, for their support and help on this unexpected path I decided to follow.

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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESUMEN

El desarrollo de sondeos sinópticos del cielo en los últimos años ha generado cantidades masivas de datos. Su análisis requiere por lo tanto recursos que superan las capacidades humanas. Por esta razón, las técnicas de aprendizaje de máquina se han vuelto esenciales para procesar esta información y extraer todo el conocimiento posible. En este trabajo se presenta una nueva metodología automática para descubrir objetos anómalos en grandes catálogos astronómicos.

De manera de aprovechar toda la información que se tiene de estos objetos, el método propuesto se basa en un algoritmo supervisado. En particular, se entrena un clasificador *random forest* con objetos de clases conocidas y se obtienen los votos de clasificación para cada uno de ellos. En una segunda instancia, se modela la repartición de estos votos con una red de Bayes consiguiendo así su distribución conjunta. La idea tras de esto es que un objeto desconocido podrá ser detectado como anomalía en la medida que sus votos de clasificación tengan una baja probabilidad conjunta bajo este modelo.

Nuestro método es apropiado para explorar bases de datos masivas dado que el proceso de entrenamiento se realiza de forma *offline*. Testeamos nuestro algoritmo en 20 millones de curvas de luz del catálogo MACHO y generamos una lista de candidatos anómalos. Luego de realizar un análisis, los dividimos en dos clases principales de anomalías: artefactos y anomalías intrínsecas. Los artefactos se deben principalmente a variaciones de la masa de aire, cambios estacionales, mala calibración o errores instrumentales y fueron por lo tanto removidos de la lista de anomalías y agregados al set de entrenamiento. Después de re-entrenar y ejecutar nuevamente el modelo llevamos a cabo una fase de post-análisis consistente en buscar información de los candidatos en todos los catálogos públicos disponibles. Dentro de nuestra lista identificamos ciertos objetos escasos pero conocidos tales como estrellas Cefeidas, variables azules, variables cataclísmicas y fuentes de rayos X. Sin embargo, para ciertas anomalías no se encontró información adicional. Fuimos capaces de agrupar algunas de estas en nuevas clases variables. No obstante, otras, que emergieron como únicas en su comportamiento, tendrán que ser examinadas por telescopios de manera de realizar un análisis en profundidad.

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.

Our method is suitable for exploring massive datasets 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 and rerunning the model, we continued with a post analysis stage by preforming 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 few individual outliers that will be followed up in order to do a deeper analysis.

1. INTRODUCTION

1.1. Data revolution

The evolution of techniques and advances in technology that have taken place in the past few decades have led to a massive amount of information far beyond the reach of available data management tools. In fact, 90% of the data available today have been collected in the last two years; in 2012, 2.5 exabytes of information were gathered daily*. Even though this phenomenon is one of the greatest achievements of humanity it is also a double-edged sword. We must now face the challenges of storage, transfer, visualization, analysis, searching and sharing. Furthermore, this is not only affecting computer scientists, but also economists, mathematicians, astronomers, biologists and scientist in almost every area that is data driven. This has also affected the way science is conducted. In the past, the scientist identified a problem, formulated a hypothesis and then collected data in order to prove or reject the hypothesis. Today, data recollection has become the first step of the process and it is the scientist's job to ask the right questions in order to take the maximum advantage of this massive amount of data.

The need to develop the right tools for analyzing this information is consequently one of the biggest tasks we face today. It is our job as scientists to face the new age we are entering in: the big data era.

*http://www.ibm.com/big-data/us/en/

1.2. Big data era in astronomy

An astronomical survey comprises a set of images or spectra of objects taken of a particular region of the sky for a certain period of time. These objects are monitored and observed by powerful telescopes, with diameters up to 10 meters and large field of views (e.g. Pan-STARRS' field of view is 7 square-degrees). Figure 1.1 shows one image obtained with a 2.5 meters telescope from the Sloan Digital Sky Survey (SDSS) .

FIGURE 1.1. SDSS image of the Whirlpool galaxy (York et al., 2000)

Over the past few decades the development of better and more precise telescopes, CCD technology and computers has led to the proliferation of large astronomical surveys. Some

of the most important surveys include MACHO (Alcock, Allsman, Alves, Axelrod, Bennett, et al., 1997), EROS (Ansari, 2004), OGLE (Udalski et al., 1997), Pan-Starrs (Hodapp et al., 2004), SDSS (York et al., 2000) and future surveys LSST (Tyson et al., 2002) and SKA (Schaubert et al., 2003). These observations result not only in valuable information but also in an immense amount of data (going up to 150 PetaBytes) to be stored and processed: astronomy must also face the challenges of the big data era. Indeed, astronomers traditional data analysis is insufficient to deal with the data while performing the analysis by visual inspection is becoming unrealistic. Consequently, the idea of developing automatic and robust methods based on machine learning and statistics is growing strength.

After the observations are completed, an exhaustive analytical stage becomes necessary in order to study the information obtained. In this context, a scientific field has been created, known as "time domain astronomy" (TDA). Its aim is to the study and analyze astronomical objects that change over time, such as pulsating variable stars, cataclysmic and eruptive variables, quasi-stellar objects and gravitational microlensing, among others (Percy, 2007). These analyses comprises mainly characterization, classification and novelty detection of these stellar objects and events.

To perform the aforementioned, the measurements obtained from the telescopes, namely the electromagnetic radiation from astronomical objects, must be converted into manageable information. These raw data are consequently processed by several methods and techniques (Wall & Jenkins, 2012) in order to transform them into standard units of flux or intensity. The result of this conversion is a set of time series, called light-curves, for every object in the database. A light-curve is then a plot of the magnitude of brightness of a star

as a function of time (usually measured in Julian dates). It is worth noting that the y-axis is plotted in descending order of magnitude. This is because magnitude is inversely proportional to the brightness of the observation. In other words the higher the magnitude of a star, the fainter it looks on the sky (Percy, 2007). An example of a light-curve is shown in figure 1.2.

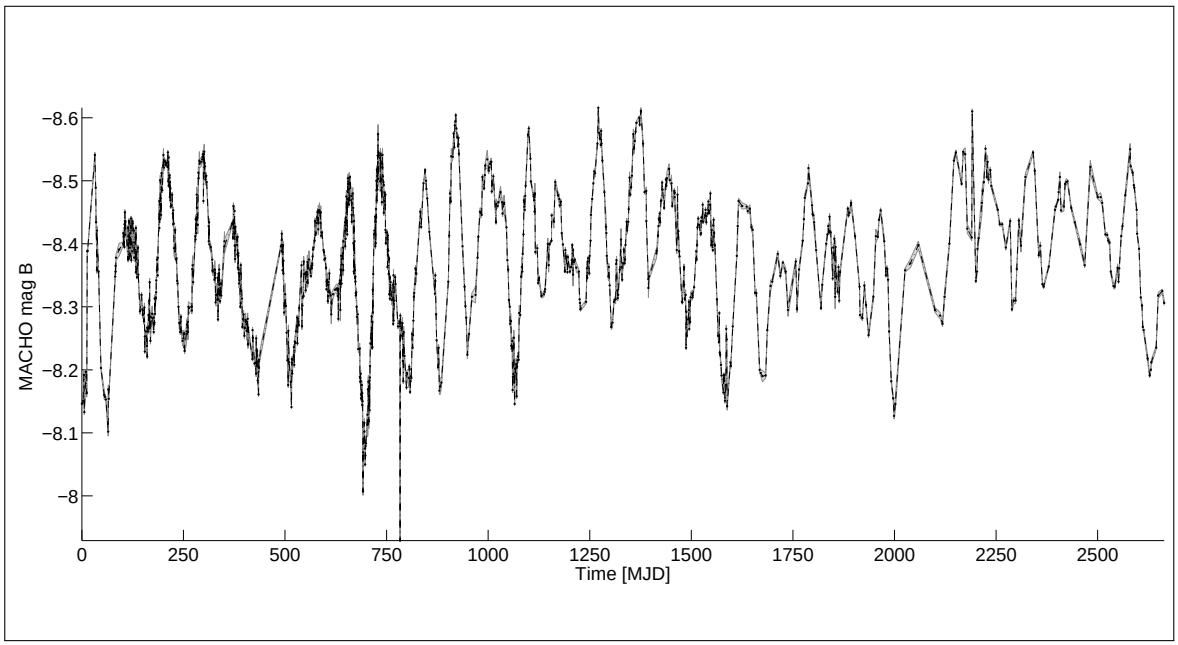

FIGURE 1.2. Light-curve example from the MACHO catalog.

1.3. Contribution of this thesis

Several important discoveries in astronomy have happened serendipitously while astronomers were examining other effects. For example, William Herschel discovered Uranus on March 13 1781(Herschel, 1857) while surveying bright stars and nearby faint stars. Similarly, Giuseppe Piazzi found the first asteroid, Ceres, on January 1 1801 (Serio et al., 2002) while compiling a catalog of stars positions. Equally unexpected, was the discovery of the

cosmic microwave background radiation (CMB) in 1965 by Arno Penzias and Robert Wilson, while testing Bell Labs horn 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 not to be behind the telescope anymore, but to be capable of selecting and making the interpretation of the increasing amount of data that technology is providing.

Outlier detection, as presented here, can guide the scientist on identifying unusual, rare or unknown types of astronomical objects or phenomena (e.g. high redshift quasars, brown dwarfs, pulsars and so on). These discoveries might be useful not only to provide new information but 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.) termed *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 thesis 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 big 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 data base 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 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 thesis we face these constraints by creating an algorithm able to efficiently deal with big data and capable of exploring the data space as exhaustively as possible. Furthermore, we address this matter from a different point of view as the one presented by He & Carbonell (2006) as "the new-class discovery challenge". Contrary to unsupervised methods, it relays 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 whole dataset

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. 1.3. The blue and green points represent instances in a two dimensional feature space from known class 1 and class 2 respectively. The shaded areas represent the boundaries learned from a classifier. The grey 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 naivest 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 as it 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, the way the classifier is confused when it comes to voting.

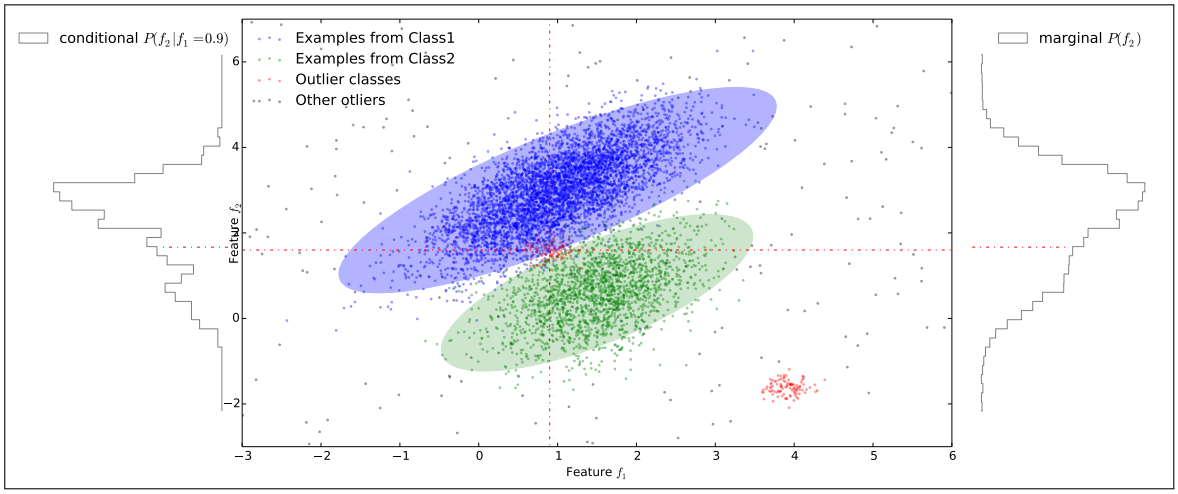

FIGURE 1.3. Simple illustration of the method.

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 way 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 a) we remove them systematically from our data set and b) create classes of spurious objects that we add to our training set. We then re-run the algorithm and obtain new candidates. These steps are repeated until obtaining no 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 consequently to give them an astronomical interpretation. Finally we cross-match the most interesting outliers with all publicly available catalogs with the aim of verifying if there is any additional information about them. In particular we are interested in knowing if 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 catalogation (Pichara & Protopapas, 2013; Kim et al., 2014). Starting with the RF output, we construct a Bayesian Network (BN) with the purpose of extracting the classifications patterns which we use for our final score of outlier detection.

1.4. Overview of this Thesis

This thesis is based on the paper *Supervised detection of anomalous light-curves in massive astronomical catalogs* by the authors Isadora Nun, Karim Pichara, Pavlos Protopapas and Dae-Won Kim that was submitted to The Astrophysical Journal on March, 2014 and it is now in the review process (Nun et al., 2014).

The thesis is organized as follows: In section 2 we detail the background theory including the basic blocks of RF and BN. Section 3 is devoted to methods related to anomaly detections in machine learning and astronomy. Our approach and the pipeline followed in this thesis are shown in section 4. Section 5 contains the information about the data used in this work and presents the results of the performed tests and the experiments with real data,

including re-training, elimination of artifacts and the post analysis process. Conclusions follow in section 6.

2. BACKGROUND

2.1. Astronomical background

In the following section we present some of the objects and phenomena that can be detected through sky surveys. In particular we focus on objects that have brightness fluctuations over time and therefore vary in the optical spectrum. This astronomical area is the main scope of this thesis.

2.1.1. Variable Stars

An important field of contemporary astronomy is the study of variable stars. Studying the variability of those objects provide astronomers with additional parameters such as time scales and amplitudes . These parameters can be useful not only to infer physical characteristics of the stars such as radius, mass and luminosity, but also to study stellar evolution and the distribution and size of our Universe (Percy, 2007; Huijse et al., 2014).

Variable stars can be divided into two main classes depending on the phenomena responsible of their brightness fluctuations. Intrinsic variable stars have internal physical changes that explain these variations while extrinsic variable stars are those in which the light output changes due to some process external to the star itself.

The tree diagram in figure 2.1 shows the classification of variable stars and a brief description of each class is presented in table 2.1.

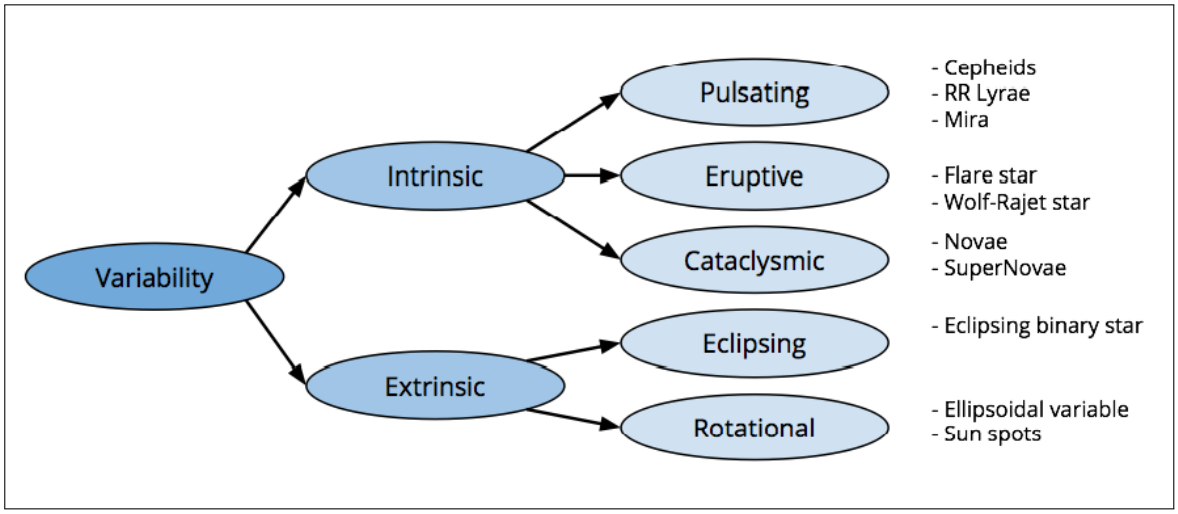

FIGURE 2.1. Variable star topological classification as presented in Huijse et al. (2014)

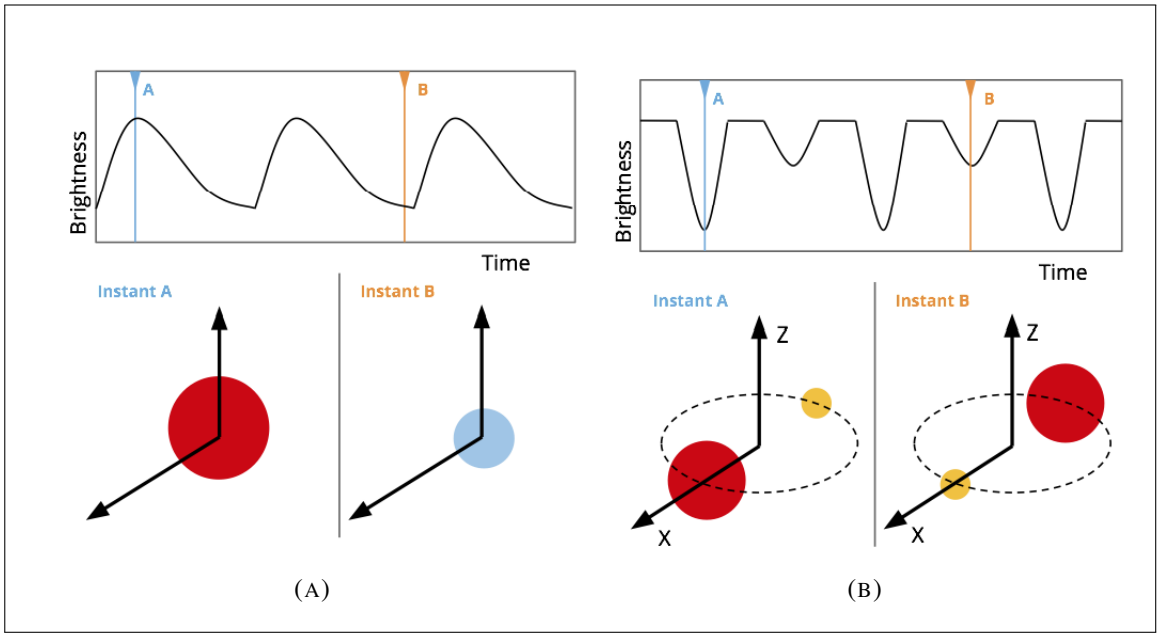

FIGURE 2.2. (a) Light curve of a pulsating variable star (upper left panel), such as a Cepheid or RR Lyrae. The star pulsates periodically changing in size, temperature and brightness which is reflected on its light curve. (b) Light curve of eclipsing binary star (upper right panel). The lower panels show the geometry of the binary system at the instants when the eclipses occur. The periodic pattern in the light curve is observed because the Earth (X axis) is aligned with the orbital plane of the system (Z axis) (Huijse et al., 2014)

TABLE 2.1. Examples of variable stars as presented in Huijse et al. (2014)

Object name	Variability	Description
Cepheid	Intrinsic	Radially pulsating supergiant star. It expands and contract periodically changing its size, temperature and brightness. The period ranges between 1 and 100 days. Fig. 2.2a shows a diagram of a pulsating variable.
RR Lyrae	Intrinsic	Radially pulsating star. Older, with a lower mass, and less luminous than Cepheids. The period ranges from 0.3 to 1.2 days.
Mira	Intrinsic	Pulsating red giant star. These variables are a thousand times brighter than our sun and have periods longer that 80-days. The Miras and other giant stars belong to the Long Period Variable (LPV) class.
Flare star	Intrinsic	Eruptive variable star. Material is violently ejected from their corona (the most outer region of a star) in a non-periodic basis. The brightness increases during this event.
Nova	Intrinsic	Cataclysmic binary star system with very close components. The hotter component "steals" material from the cooler component which then becomes the catalyst of the explosive reaction. The brightness of the Novae increases in $\sim$10 orders of magnitude and then gradually returns to its former brightness.
SuperNova	Intrinsic	The final explosion of a massive star. During this transient event, the brightness of the system increases by $\sim$20 orders of magnitude. The material ejected from the explosion forms a gas cloud (the basis from which new stars will be formed) and the remnant of the star will evolve into either a neutron star or a black hole depending on the its original mass.
Eclipsing Binary	Extrinsic	Binary star system with orbital plane alligned with Earth. The mutual eclipses appear as periodic brightness drops in the light curve. A diagram showing the geometrical configuration of the system is shown in Fig. 2.2b.
Pulsar	Extrinsic	Highly magnetized and dense neutron stars with fast rotational speeds. Pulsars emit electromagnetic radiation with periods that range between 1ms and 10s. This emissions can be detected if the emission axis of the pulsar is aligned with Earth.

2.1.2. Variable non-stellar phenomena

Besides variable stars, there are other phenomena known by their brightness fluctuations. Table 2.2 summarizes some of them.

TABLE 2.2. Examples of non-stellar variable sources as presented in Huijse et al. (2014)

Name	Description
Gravitational lensing	An increase of several orders of magnitude in the observed brightness of a star due to a massive dark object passing in front of it and acting as a lens. The dark object bends the light of the source. If the dark object is of planetary size the effect is called microlensing.
Active Galatic Nuclei	A compact region in the center of a galaxy characterized by their variable, strong and broad electromagnetic emissions. The most studied AGN is the quasar (quasi stellar radio sources). Quasars are one of the most luminous objects in the sky and are characterized by their strong stochastic variability across wavelengths and on timescales.
Transiting extrasolar planet	Planet outside our solar system, orbiting a star different than our sun. If the orbital plane of the exoplanet is aligned with the Earth, periodic drops will appear in the light curve of its main star. Small planets may induce a shallow minimum that needs to be discriminated against the noise of the light curve.

2.2. Machine learning background

Machine learning is an area of artificial intelligence that provides computers with the ability to learn and make predictions from the data without being explicitly programmed. There are two basic types of algorithms in this field: supervised and unsupervised.

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

2.2.1. Supervised learning algorithms

In the supervised learning algorithm case, there is a labeled set of input-output pairs (the training set) from which we infer or learn a general mapping from inputs x (the features describing the object) to outputs y (the object label). For example, consider a set of input

data of temperature measures from several days and a labeled output corresponding to each day weather (rainy, cloudy and sunny). The training data is used to infer the values of relevant parameters which are then utilized to make forecasts for a new data point (the test set). Below, we summarize some of the basic supervised machine learning algorithms.

- Decision trees (Breiman et al., 1984): flowchart-like structure where each internal node denotes a test on an attribute, each branch corresponds to an outcome of the test and each external node or leaf denotes a class prediction or a decision. At each node, the algorithm chooses the "best" attribute by following a particular set of criteria (for example information gain) in order to partition the data into individual classes. Figure 2.3 shows one example of such model proposed in Quinlan (1986) for a problem that involves weather. To simplify the case it is assumed that there are only two classes denoted P and N, referred to as positive and negative instances, respectively.

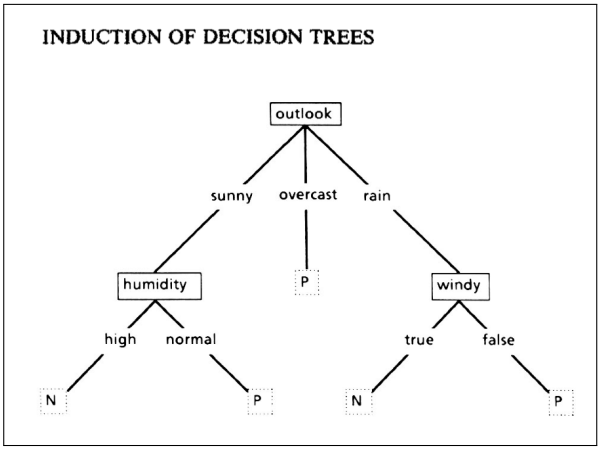

FIGURE 2.3. Decision tree example. Only two classes denoted P and N are considered, referred to as positive and negative instances, respectively. Retrieved from http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf

- Support vector machines (SVM) (Boser et al., 1992): two-class classification model that defines a linear hyperplane or set of hyperplanes in a high dimensional space that separate the data. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (known as margin), since in general a larger margin implies a lower generalization error of the classifier. Figure 2.4 illustrates SVM algorithm.

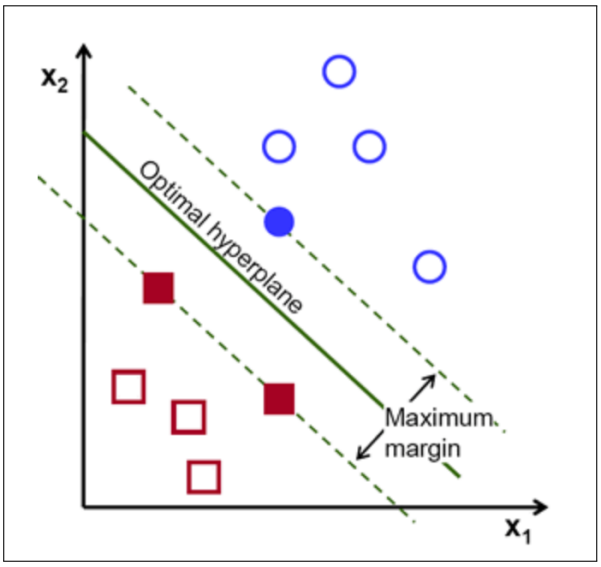

FIGURE 2.4. Example of SVM: the hyperplane maximizes the margin of the training data. Retrieved from http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

- Artificial neural networks (ANN) (Rumelhart et al., 1988): model inspired by the central nervous system. It is composed of interconnected units that serve as "neurons" on a system with three basic elements. First, the synapses of the biological neurons are modeled as weights and therefore they reflect the strength of the connections. The following components of the model represent the actual activity of the neuron cell: all weighted inputs are summed. This activity is

referred to as a linear combination. Finally, an activation function normalizes the amplitude of the output (between 0 and 1, or -1 and 1 for example). A representation of this process is illustrated in figure 2.5.

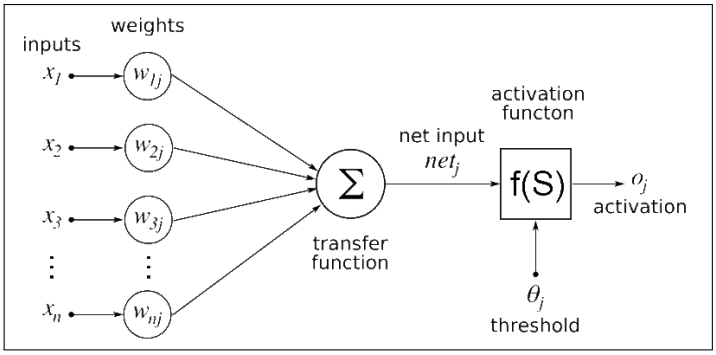

FIGURE 2.5. ANN process: interconnected neurons, learning process for updating the weights of the interconnections and activation function that converts a neuron's weighted input to its output activation. Retrieved from http://www.durofy.com/machine-learning-introduction-to-the-artificial-neural-network/

2.2.2. Unsupervised learning algorithms

In unsupervised learning algorithms in contrast, there is no labeled training set. A simple set of input data is available from which we wish to find interesting patterns. Clustering algorithms (Hartigan, 1975) are the main methods of unsupervised learning since their aim is to find meaningful groups of similar members without necessarily having any predefined classes. A good clustering is achieved when the intra-cluster similarity is maximal (low distances between objects of the same cluster) and the inter-cluster similarity is minimal (high distances between objects of different clusters) as shown in figure 2.6. Some of the clustering algorithms are:

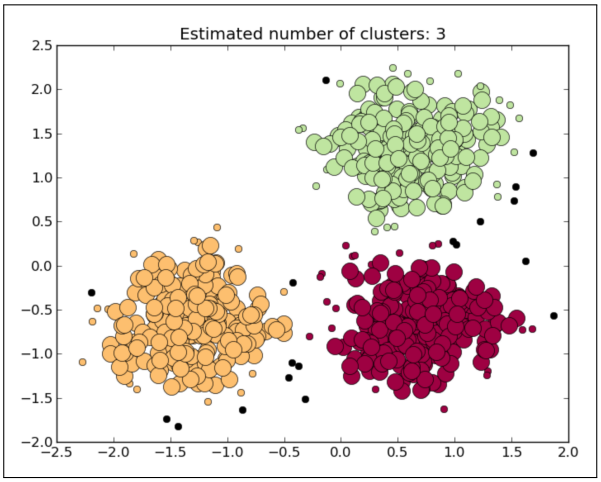

FIGURE 2.6. Example of clustering: each color represents a different cluster. Retrieved from http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html

- K-means (Hartigan & Wong, 1979): iterative method that aims to classify the data into k different clusters (user specified parameter). The main idea is to define k centers points or centroids, one for each cluster. On the first iteration the centroids are randomly selected and each point is assigned to its closest centroid based on a distance metric. The new centroids are then recalculated as the mean of all the points assigned to each cluster and the objects are reassigned. The iterations continue until the assignment is stable: the clusters formed in the current round are the same as those formed in the previous round.
- DBSCAN (Ester et al., 1996): method that finds objects with dense neighborhoods, also known as core objects. A point density can be measured as the number of instances close to it in a defined ϵ radius (user specified parameter). The algorithm connects then core objects and their neighborhoods to form dense regions as clusters. To determine if a neighborhood is dense, DBSCAN uses

an other user-specified parameter, *MinPts*, which specifies the density threshold of dense regions. Therefore, an object is considered as a core object if its ϵ-neighborhood contains at least *MinPts* objects.

- EM-Gaussian mixture of models (Reynolds, 2009): soft clustering algorithm where data points are assigned to k clusters with certain probabilities. The method assumes that each cluster contains data generated from a Gaussian distribution with parameters (μ_j, σ_j) (the mean and standard deviation of each j distribution) that need to be learned. For each data point, the probability of belonging to each cluster is not known (not observed) - called latent variables- thus it also needs to be learned. The algorithm starts from some initial estimate of the parameters (e.g., random), and then proceeds to iteratively update until convergence is detected. The algorithm steps are the following: a) Randomly initialize the Gaussian parameters and assume equal probability amongst classes b) E-step: compute the probability of each data point of belonging to each cluster for all data points and all clusters given the parameters at this iteration c) M-step: recompute the parameters of each Gaussian.

2.2.3. Model selection

Generally, the first task in tackling any real world problem in machine learning is to select a model. Even if the main interest of most of the problems is to obtain a particular output variable, the choice of the relevant inputs to perform this inference can be highly determinant. In addition, deciding how exactly the output variables are related to a given

set of input variables, and how the latter are related to each other can be very tricky, and in many cases may have multiple answers. Nevertheless, there is one principle to follow when choosing the model. Occam's razor, also known as the parsimony principle, establishes that the best model is the simplest one that adequately describes the data. Oftentimes, it is possible to find a more complex model that performs better when training, but then it performs worse on the test set (out-of-sample), it will perform worse than a simpler model. This problem is described by Dietterich (1995) as overfitting.

2.2.4. Model evaluation

Besides avoiding the overfitting obstacle, the modeling goal should be to minimize the generalization rate: the expected value of misclassification rate when averaged over future (test) data. In other words, it is the ability to perform accurately on new, unseen examples or tasks after having experienced a training data set. Without having access to the future data, the analysis of the misclassification rate (the percent of the training data misclassified by the model) is used to calibrate the model for high future performance. However, due to the overfitting problem, a model that minimizes the misclassification rate will not always perform optimally out-of-sample. There are many approaches that are commonly used to avoid this constraint. One popular example, used in this thesis, is the K-fold cross validation technique. The idea is to divide the training data into k groups. The model is then trained with $k-1$ groups and tested with the k-th one in a round robin fashion. Once this process in performed, the model that minimizes the average prediction error from all groups is chosen. Figure 2.7 shows a pictorial representation of this process.

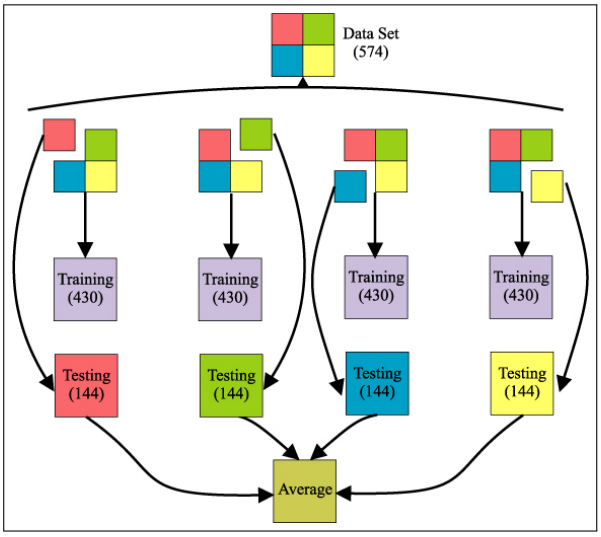

FIGURE 2.7. Cross validation for a 574 objects training set. The data is divided into k groups. The model is then trained with $k-1$ groups and tested with the k-th one. The model that minimizes the average prediction error from all groups is chosen. Retrieved from Ruiz et al. (2013)

It is worth noting that even in cases where it is known with certainty that the data is truly high dimensional, there might be only several latent factors which describe most of the output variability. Therefore complexity reduction could be achieved with minimal loss of accuracy. Since more complex models are often associated with great computational costs, it is generally useful to reduce dimensionality whenever is possible. Model exploration and pre-processing is therefore always important and should always be performed before making the inference.

2.2.5. Random forest

Random forest 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 decisions 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 a RF model given some training data is as follows:

- Let R be the number of trees in the forest (a user defined parameter) and $|F|$ be the number of features describing the data.
- 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 than the training set but some of the examples are selected more than once, given that the samples are taken with replacement.
- For each of the R sets, train a decision tree using at each node a feature picked 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, training each individual tree 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 only a subset of randomly selected features is used while building each tree. This contributes

to the reduction of the dimensionality and has been shown to significantly improve the RF accuracy (**?**Geurts et al., 2006).

2.2.6. Bayesian Networks

A Bayesian network 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 BN is to estimate joint probability density functions (PDF). 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:

$$P(v_1, \ldots, v_k) = \prod_{i=1}^{k} P(v_i | Pa_{BN}(v_i)) \tag{2.1}$$

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 a) to learn the set of edges ε, or in other words the BN structure, and b) to learn the conditional probabilities $P(v_i | Pa_{BN}(v_i))$.

2.2.7. Learning the edges of the BN

Recall that a BN is a directed acyclic graph where each node represents a random variable. In our case the random variables we are modelling 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 each node, greedily adding a new candidate parent and checking if the new addition creates a better network score or not. In case 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 apply the same factorization imposed by the structure to the data and use multinomial distributions over each factor. How exactly score is

assigned to a given structure is well described in the original work (Cooper & Herskovits, 1992; Pichara & Protopapas, 2013).

2.2.8. 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 discretise 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 do the discretisation, the data is divided into a set of bins, thus every data value which 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 2.3.

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

$$(N_{bins} - 1) \times (N_{bins})^{N_{parents}} \tag{2.2}$$

TABLE 2.3. Probability of v_j given the different values of the parents, $P((v_j|v_a, v_b))$. 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.

	$v_j = 1$	$v_j = 2$
$v_a = 1, v_b = 1$	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	$1 - \theta_4$

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, lets say θ_1, is calculated as:

$$P(\theta_1|data) = \frac{P(data|\theta_1) \times P(\theta_1)}{\sum_{\theta_1} P(data|\theta_1) \times P(\theta_1)} \tag{2.3}$$

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:

$$P(data|\theta_1) = \theta_1^{N_1} \times (1 - \theta_1)^{N_2} \tag{2.4}$$

Where N_1 is the number of cases in the data where v_j's take 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 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 manner to simulate previously seen "imaginary data" in order to compensate situations of 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:

$$P(\theta_1|data) \propto \theta_1^{N_1+\alpha_1} \times (1-\theta_1)^{N_2+\alpha_2} \tag{2.5}$$

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.

3. RELATED WORK

3.1. Outlier detection in machine learning

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

In unsupervised approaches the examples given to the learner are unlabeled and consequently there is no training set in which the data is 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 data that does 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, a 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 which 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 requires to assume a distribution for data, whereas in complex cases, most of the known distributions will not fit data the way we expect.

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 lies 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 on the context of intrusion detection on networks (Yang et al., 2006; Son et al., 2009). Unfortunately, clustering algorithms suffer from the curse of 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 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 kind of approaches is that they use heuristics to find relevant

subspaces and those heauristics may ignore combinations of spaces where anomalies could also lie.

Finally, proximity-based methods follow the intuition that anomalies are records with less 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 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 hyper-sphere that is centered at the data instance and contains k (nearest) neighbors. Papadimitriou et al. (2003) propose a variant of the LOF called 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 micro clusters 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 micro clusters detected. Unfortunately, density-based algorithms usually are quadratic in the number of instances and thus they are not suitable for big data. Furthermore, these methods also suffer from the curse of dimensionality because of 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 trees 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 which are out of the scope of the training set. Decision trees and Neural networks are susceptible to overfitting when stopping criteria are not well determined.

3.2. Outlier detection in astronomy

Because synoptic sky surveys have significantly increased in the last decade (Keller et al., 2007; Hodapp et al., 2004; Tyson et al., 2002), astronomical anomaly detection has not been yet 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-curves 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 outliers light-curves in catalogs of periodic variable stars. To this end, they use cross-correlation as measure of similarity between two individual light-curves and then classify light-curves with 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 bases features. Quite 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 it 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 Dirichlet genre model (DGM), which assume 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 focus on the distribution of objects in the group. One of the main drawbacks of this method is that the inference stage considers a non convex problem and is consequently restricted to the limitation of variational approximations.

Finally, Henrion et al. (2013) propose CASOS, an algorithm to detect outliers in datasets 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 allusion 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 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.

4. DESCRIPTION OF THE PROPOSED MODEL

In the next section we detail our work and methodology. For illustration, we present in Figure 4.1 a pictorial representation of our algorithm and its two main stages: the training stage and the outlier detection stage. 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 an outlier.

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 5.1, we give details about the classes and statistical descriptors we used.

We train a RF classifier and obtain voted labels for each element using the set S. More precisely, we perform a T-fold cross validation. To do so, the original data is randomly partitioned into T equal size subsamples or *folds*. For each iteration, we train a RF with $T-1$ folds and we use that RF to predict the label for the fold that the RF did not see in the training process. Repeating this process for each fold, we end up with predicted labels for

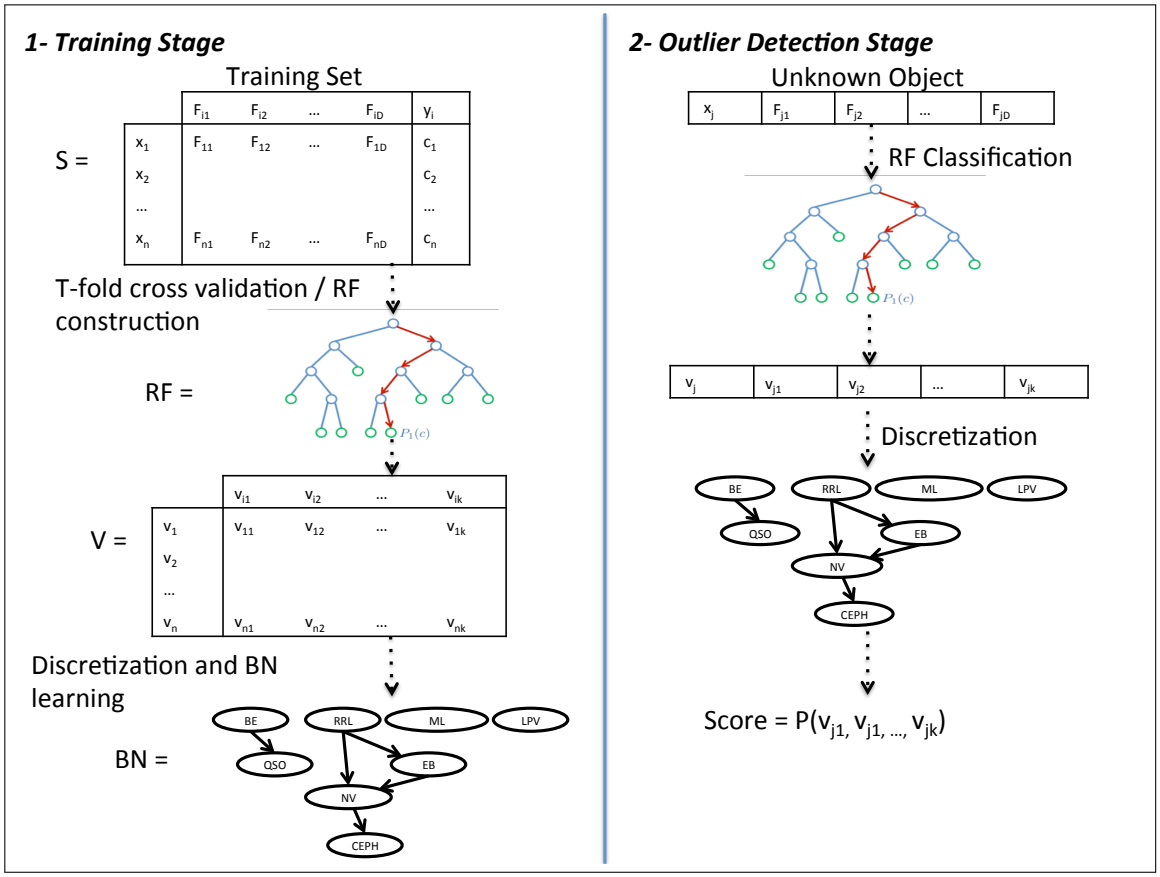

FIGURE 4.1. Algorithm illustration. The left panel shows the training stage of our method and the right panel presents the anomaly detection process. In the training stage, we start with a training set followed by the training of the RF, discretization of the probabilities and finally the construction 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 an outlier.

each element in the training set S. Each prediction obtained from the RF come as a vector $\{v_{i1}, \ldots, v_{ik}\}$ where each $v_{ij} \in [0, 1],\ j \in [1 \ldots k]$, tell us the probability that the element x_i belongs to the class y_j, $\sum_{j=1}^{k} v_{ij} = 1,\ \forall i \in [1 \ldots n]$.

In our experiments we use $T = 1000$ folds, 20 bins for the discretization and a maximum of three parents.

At the end of the cross validation process, we end with a new dataset $V = \{v_1, \ldots, v_n\}$ where each $v_i = \{v_{i1}, \ldots, v_{ik}\}$. This dataset gives us information about how the RF votes among objects that belong to each of the known variability classes. We want to use this dataset to decide if an unlabeled object belongs to an unknown variability class or not, simply by comparing the RF votes of this unlabelled object with the "usual" votes of the RF obtained from the dataset V. If the voting vector for the unlabeled object is too different from the voting vectors stored in the dataset V, we flag it as an outlier. To do this comparison, we learn the joint probability distribution over the dataset 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, $P(v_1, \ldots, v_k)$.

In section 2.2.8 we mentioned the necessity of a prior in order to include all the possible cases in our model. To chose 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 bins) and the maximum number of parents a node can have on the BN. Given that the minimum number of parents is zero and the maximum is three, a reasonable number for α

is four. We also empirically tested with different values of α and found that the results are not sensitive to the choice of α.

5. DATA, EXPERIMENTAL RESULTS AND ANALYSIS

5.1. MACHO catalog

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

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

5.1.1. Training set

The training set is composed of a subset of 5646 labeled observations from the MACHO catalog (Kim et al., 2011)*. The constitution of the training set is presented in Table 1 and a representative example of each class light-curve is shown in Figure 5.1.

The catalog comprises several sources from MACHO variable studies (Alcock et al., 1996; Alcock, Allsman, Alves, Axelrod, Becker, et al., 1997c,d; Alcock et al., 1999), the MACHO microlensing studies (Alcock, Allsman, Alves, Axelrod, Becker, et al., 1997a;

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

TABLE 5.1. Training Set Composition

	Class	Number of objects
1	Non variable	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

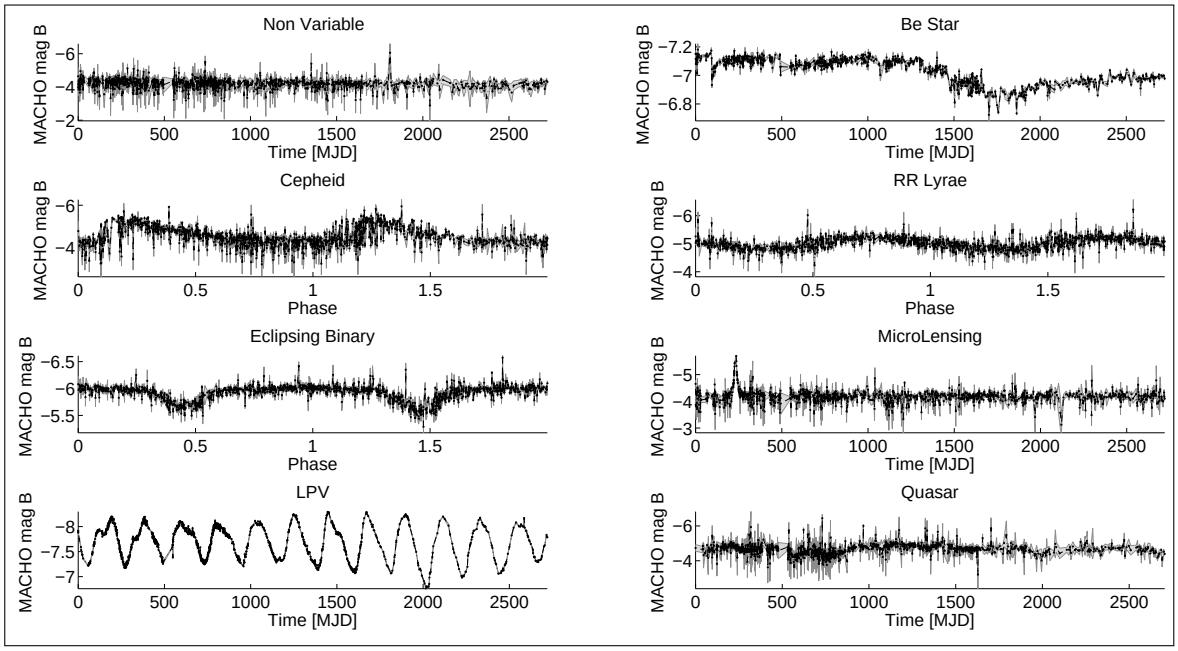

FIGURE 5.1. Example light-curves of each class in the MACHO training set. The *x*-axis is the modified Julian Date (MJD), and the *y*-axis is the MACHO B-magnitude. Note that Cepheid, RR lyrae and Eclipsing Binary light-curves are folded since they are periodic.

Alcock, Allsman, Alves, Axelrod, Bennett, et al., 1997; Alcock, Allsman, Alves, Axelrod, Becker, et al., 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 Geha, M. The non-variables were randomly chosen from

the MACHO LMC database, and any previously known MACHO variables were removed from the non-variable set.

5.2. Results

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

5.2.1. 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 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. For every test we ran a 10 fold cross validation. The RF considered 500 trees, with $F' = \sqrt{F}$ 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 5.2. By using a color scale, we also show the average distribution of the votes among the different classes. For example, when the RF is classifying a RR Lyrae it doubts mainly between non variables, 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 5.3. When the light-curve to be classified is from RR Lyrae class the voting vector will present high values for eclipsing binaries and Cepheids classes, therefore RR Lyrae node is a child node of these other two classes.

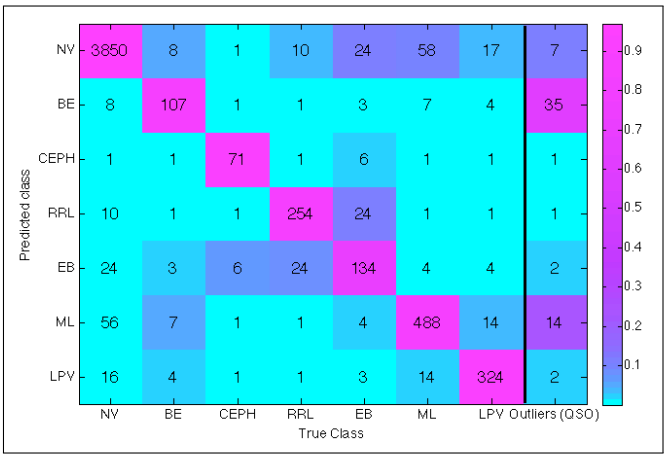

FIGURE 5.2. The color scale represents the RF votes distribution for the predicted classes in the MACHO training set (NV: Non variable, BE: BE stars, CEPH: Cepheid, RRL: RR lyrae, EB: Eclipsing Binaries, ML: Microlensing, LPV: Long Period Variable) during the cross validation phase and for the test class (QSO: Quasars) during the testing phase. The values displayed in each square represent the average number of objects that were voted in the correspondent class. The x-axis represents the object's true class and the y-axis represents the predicted class.

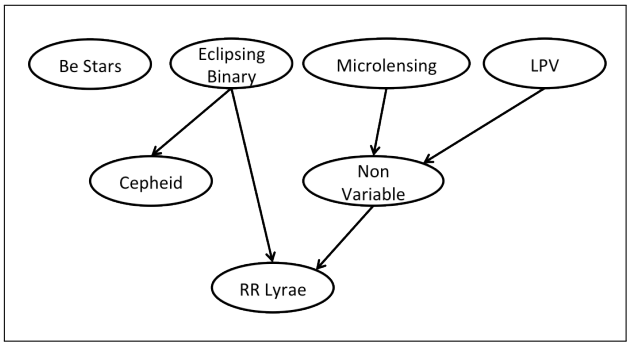

FIGURE 5.3. BN structure for the performance test. RR Lyrae node is a child node of Cepheid and non variable nodes meaning that, when the light-curve to be classified is from RR Lyrae class, the voting vector will present high values in these other two classes. On the other hand, Be stars node is independent of the other classes, as expected.

After training the algorithm, we obtained the joint probability of every object in the training set, quasars included. Figure 5.4 shows how "known" classes present a high joint probability while outliers (quasars) have the lowest values. Finally, the top left panel of Figure 5.5 represents our algorithm performance, comparing the imputed outliers (quasars) positions in the top outliers 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 panel in Figure 5.5.

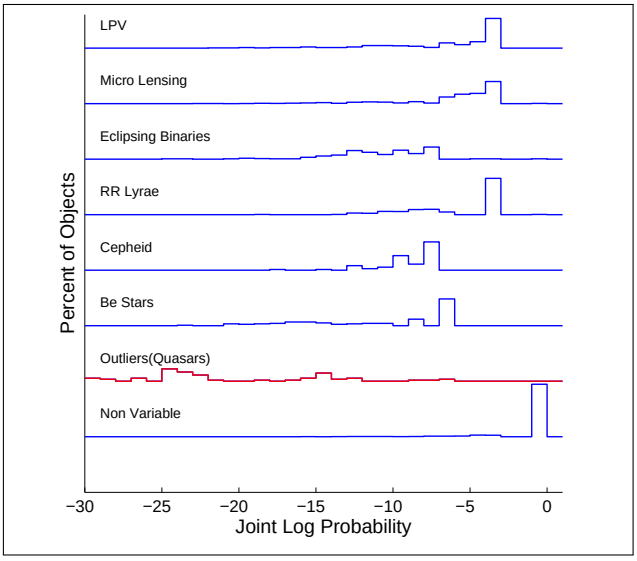

FIGURE 5.4. Joint log probability distribution for each class in the training set (blue lines) and for the outlier class (red line)

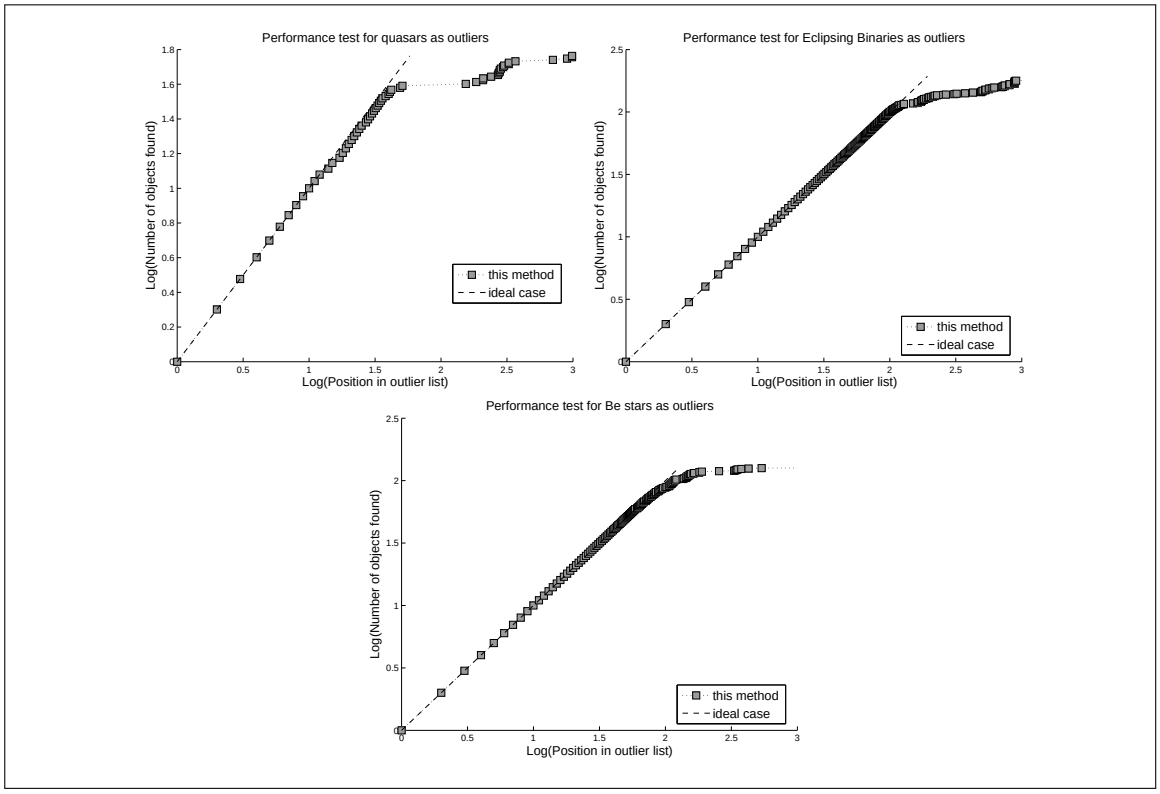

FIGURE 5.5. Performance test results for Quasars, Eclipsing binaries and Be stars as outliers. The dashed line represents the ideal result, where the class left out use the top positions in the outlier list. Grey squares shows the actual obtained positions.

5.2.2. Running on the whole dataset

Once we tested the accuracy of our method, we trained a RF 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 of 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 run the cross validation and

learn the BN structure and parameters. After training the model, performing the inference for a light-curve takes a fraction of second and it is easily parallelizable.

5.2.2.1. Removal of spurious outliers

Figure 5.6 shows some of the outliers we obtained from this first iteration. The top left and right outliers in Figure 5.6 are characterized by having one day period, 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 kind of artifacts like the outlier in Figure 5.6 bottom left panel, which is obviously due to some instrumentation problems - this behavior at the beginning of the light-curve appeared in many light-curves.

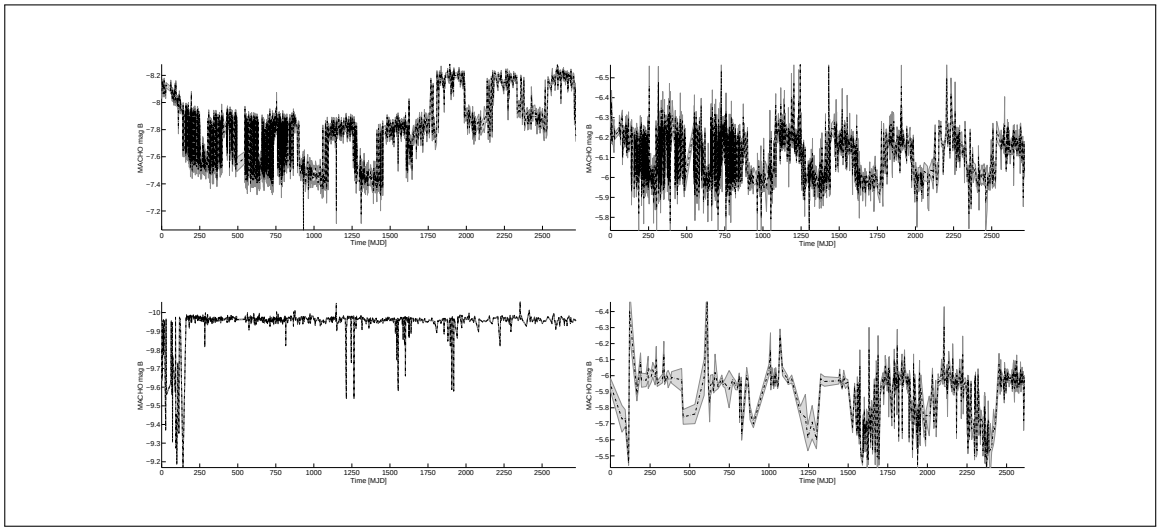

FIGURE 5.6. Top left panel one day period artifact MACHO_77.7187.271, top right panel one day period artifact MACHO_79.4780.358, bottom left panel sampling artifact MACHO_5.5010.986 and bottom right panel 370 days period MACHO_49.5899.715.

In order to remove the spurious outliers we do the following steps:

(i) Filter all outlier candidates that have periods very close to sidereal day or a year. There is no doubt that those light-curves exhibit strange behavior due to variable seeing conditions during the night or seasonal aliases.

(ii) We run the whole analysis in the MACHO red non standard 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 it is removed from the candidate list.

(iii) We visually inspect all candidates and group those that are obviously spurious, like the examples in Figure 5.6, into groups of similar shapes and behaviors. We add these new classes to the training set, re-train and then we predict outliers again as explained above.

(iv) 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.

5.3. Post analysis

As a fist step, we visually inspected all the candidates starting from the top of the list ("strongest" outliers) moving our way to the "weakest" outliers. We determined that about 4,000 candidates was a good number of candidates to start. Candidates beyond this point

either were not showing any significant variation or had low signal-to-noise ratio (SNR), and therefore 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 5.2 summarizes all the catalogs used in the analysis and the resulting cross-matched numbers ($N_{\mathrm{x-matched}}$).

TABLE 5.2. 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[†]
XMM-Newton	Watson et al. (2009)	262,902	13
ROSAT All-Sky Bright Source Catalogue (1RXS)	Voges et al. (1999a)	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

The fact that some of the candidates appear in catalogs of known objects imply that we either have false positives in our results or there are misclassifications in the catalogs. Possible reasons for false positives are:

(i) Known classes are not included in our original training set. This may be the case when a new class of objects was discovered and published after the training set was constructed.

(ii) The training set is not complete for some of these classes. For example, we find that a number of outliers turned out to be eclipsing binaries (29), Cepheids (8) or RR Lyrae (8). However these classes were used in the training set as indicated in Table 5.1. Nevertheless, as in the case of MACHO_25.4201.54 which is supposedly a Cepheid, the period in our catalog is 59.08 days, which is longer than the Cepheid periods in the training set ($1 \text{ day} < P < 50 \text{ days}$). Such objects could be identified as 'rare' Cepheids or alternatively our training set could be incomplete with respect to Cepheids.

(iii) 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 would always be present. These errors will hopefully present themselves as outliers in our final analysis.

(iv) The features considered in this work and the features used by the other catalogs are not the same. For example, we find that the period of MACHO_77.7428.190 is 906.3559 days, while in Soszynski et al. (2008) is 0.2843359 days. Because of this, this light-curve does not appear to be an RRL, in our model, however all other features indicate that it is an RRL and therefore it is identified as an outlier. Dealing with feature uncertainties is a topic of a future work. It is well known

that less confident features produce low quality classification and false-positives in our outlier predictions.

(v) The SNR of the light-curves is survey dependent and therefore features that are dependent on the actual amplitude of the variability will be different from catalog to catalog. For example if a catalog is compiled using a survey that is more sensitive than our survey, then the fainter objects are indistinguishable from the non-variables. Moreover, as described above, low signal-to-noise ratio light-curves will result into uncertain features and therefore higher probability of being false-positive.

Almost all of these reasons can be attributed to the lack of a perfect training set. Because our method is based on a supervised classification, the results heavily depend on the choice of these representative objects. In the ideal scenario, one would compile a training set that contains every possible known objects. In our case, we started with a trustworthy training set and we were aware that some of the known objects were not included. This served as a blind test since some of these types were never presented to the method, never trained with them and therefore they should have been discovered by our method. Indeed 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 5.12 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 top most interesting subgroups. Some of them are known, but are rare objects while others do not obviously belong to any known class of objects.

(i) 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 5.7 shows one of these examples.

(ii) Cataclysmic Variables (CV): Another interesting group of outliers are CVs or novae or novae-like looking objects. Because there are no unified variability characteristics, this group was not included in the training set and therefore few CVs are in our candidate list. These objects can increase more than 20 magnitudes, becoming approximately 10^8 times brighter. Novae and Recurrent Novae

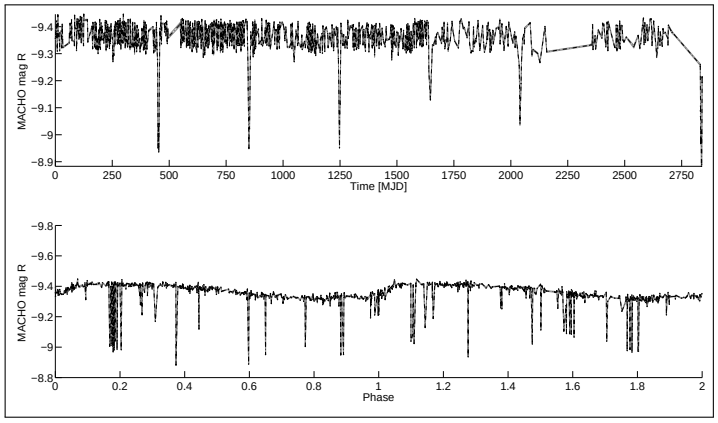

FIGURE 5.7. Top panel Eclipsing Cepheid MACHO_6.6454.5 and bottom panel its folded light-curve.

are close binary systems that are variable due 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 a subject of an extensive research, and recently the interests focused on superluminous SNe (Quimby et al., 2011). Figure 5.8 shows MACHO_77.7546.2744, one of this class example, 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).

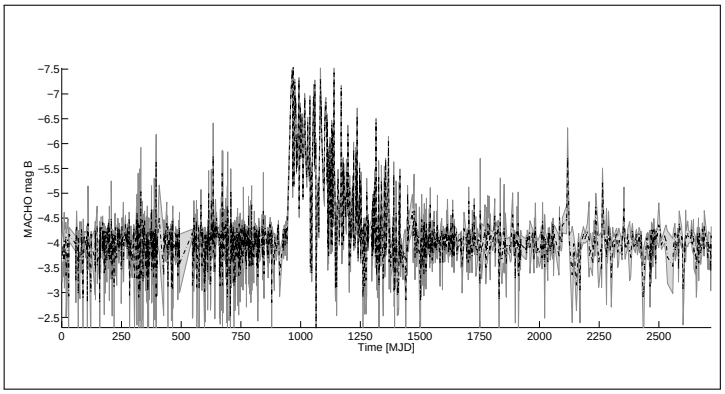

FIGURE 5.8. Nova like variable MACHO_77.7546.2744.

(iii) 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 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 light-curve is shown in Figure 5.9 and the locations of all the members in the CMD are shown in Figure 5.12.

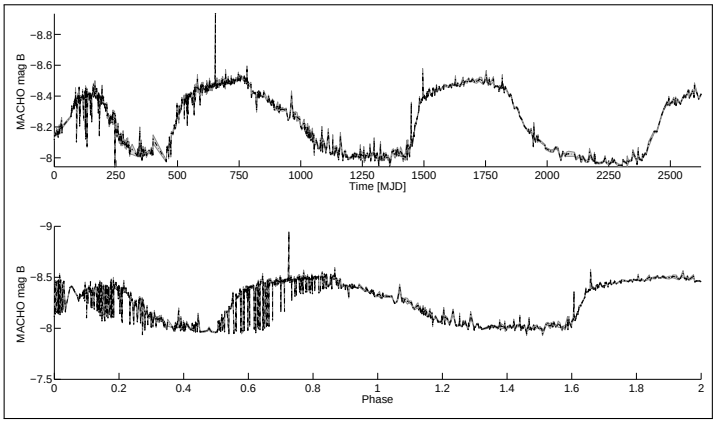

FIGURE 5.9. Blue variable MACHO_81.9727.662.

(iv) X-ray Sources: There are two sources cross-matched with the ROSAT all-sky survey bright source catalog (Voges et al., 1999b) 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 non-periodic 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 help us to understand the process of accretion and the fundamental physics of the binaries such as mass, radius, orbit, jets, etc (e.g. see van der Klis 2000; Fender et al. 2004). Figure 5.10 shows one representative example.

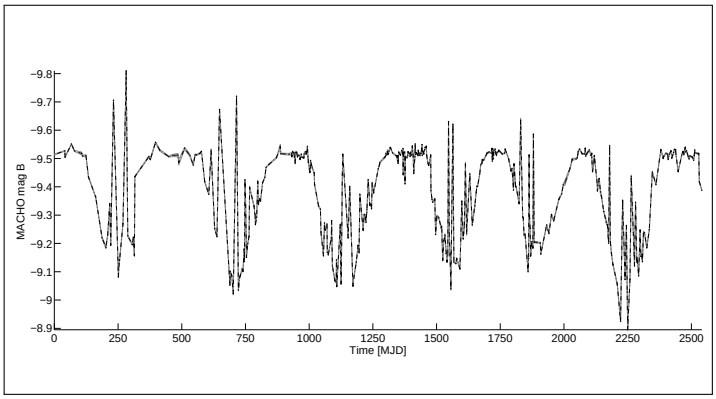

FIGURE 5.10. X-Ray binary MACHO_61.9045.32.

(v) 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 a R Coronae Borealis (RCB) 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 magnitudes in a short period(Clayton, 1996). An example of this type of light-curve is shown in Figure 5.11.

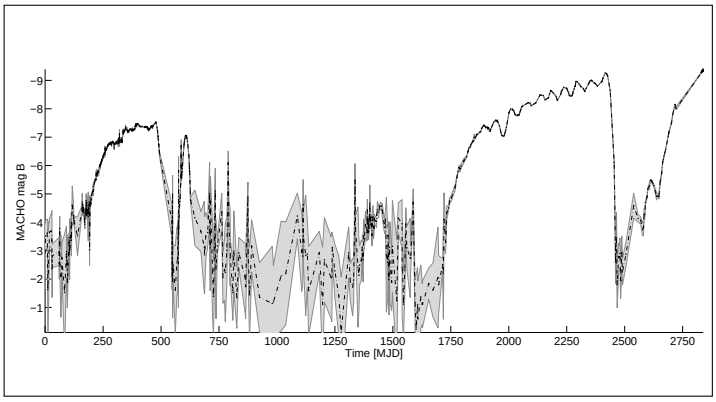

FIGURE 5.11. R Coronae Borealis MACHO_6.6696.60.

(vi) **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 and then visually inspected all the light-curves that are not in the categories mentioned above, identifying a few classes of objects and a few individual objects that could not be assigned to known classes. We show three classes and 4 individual outliers in Table 5.3 , Figures 5.13, 5.14, 5.15, 5.16, and also in the CMD in Figure 5.12.

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 (since they are located in the same field, number 82), so it is very possible 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 of 400 days but the objects were $\sim$100 arcsec apart. Since proper motion are less than few arsec/year the hypothesis was rejected. Objects of Class A are consequently good candidates to conform a new variability class.

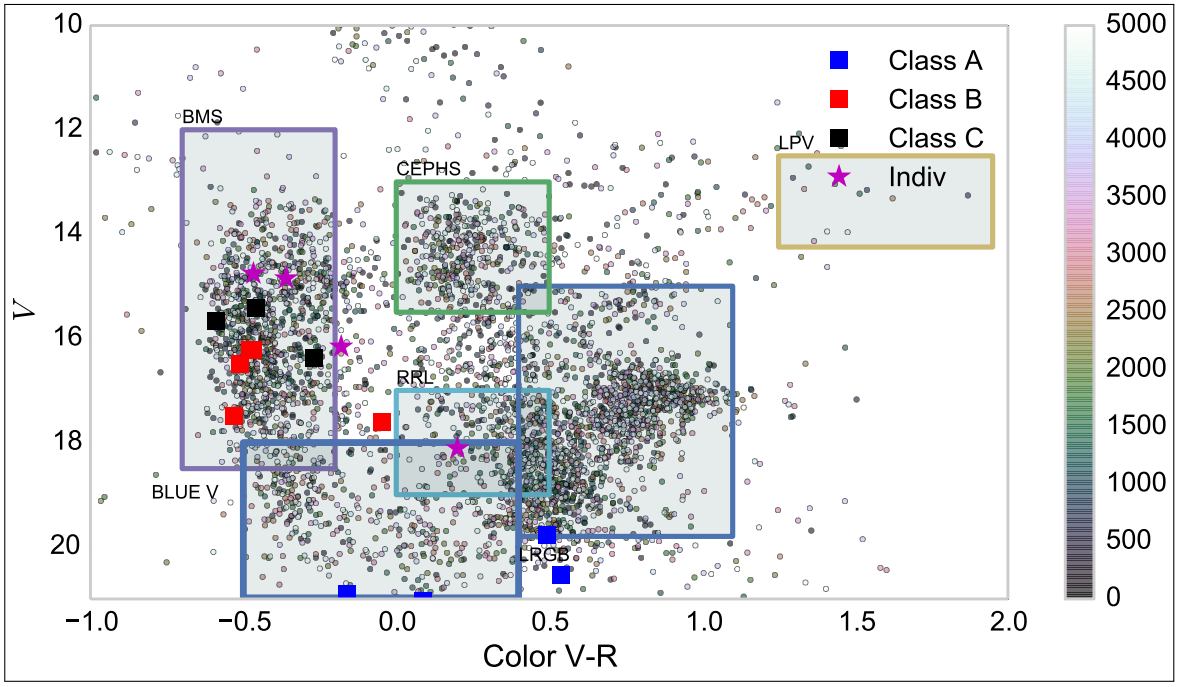

FIGURE 5.12. Color magnitude diagram of all the outliers. The outlier rank is indicated by the color of each data point. The bluer, the higher the outlier score. Black boxes mark the location of blue main sequence (BMS), lower red giant branch (LRGB), long period variables (LPV), RR Lyrae (LLR) and Cepheid (CEPH).

TABLE 5.3. The others: new variability classes and individual outliers

Class	MACHO id	RA	Dec	Period [days]	V	R	Color	SNR
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

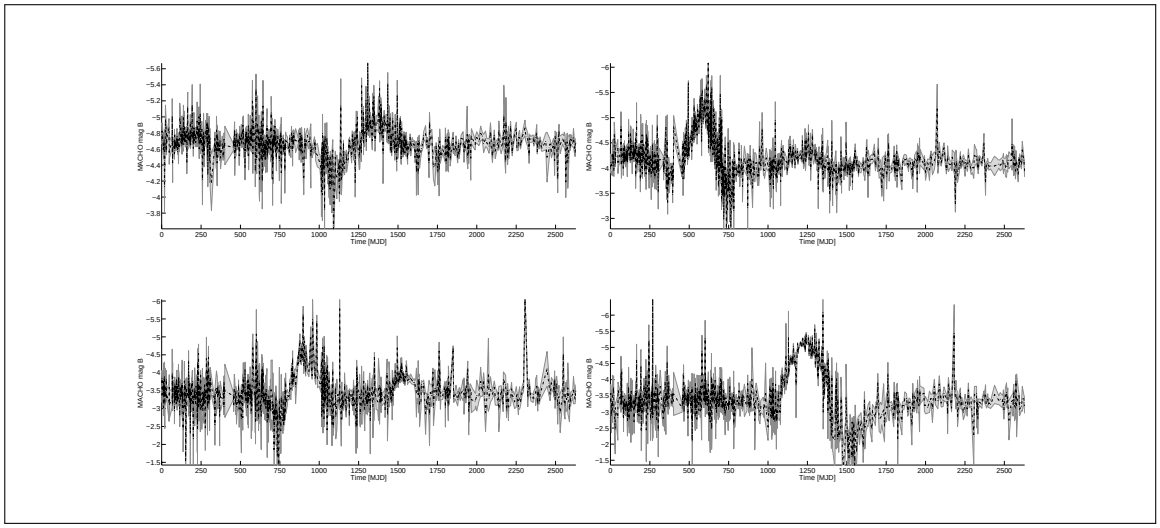

FIGURE 5.13. Top left panel Class_A MACHO_82.8887.471, top right panel Class_A MACHO_82.9009.834, bottom left panel Class_A MACHO_82.9009.1850 and bottom right panel Class_A MACHO_82.8887.2395.

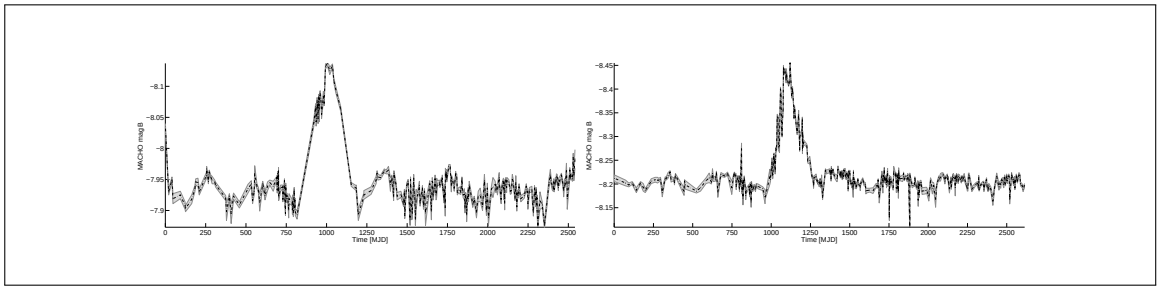

FIGURE 5.14. Left panel Class_B MACHO_56.5178.29 and right panel Class_B MACHO_44.1616.257.

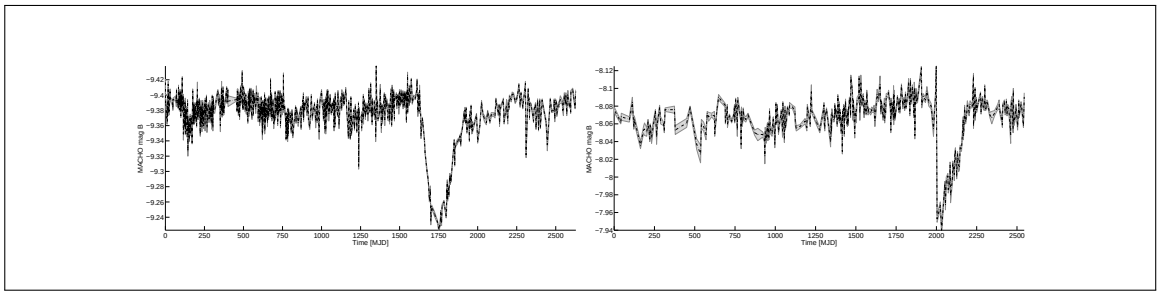

FIGURE 5.15. Left panel Class_C MACHO_82.8283.41 and right panel Class_C MACHO_62.7361.30.

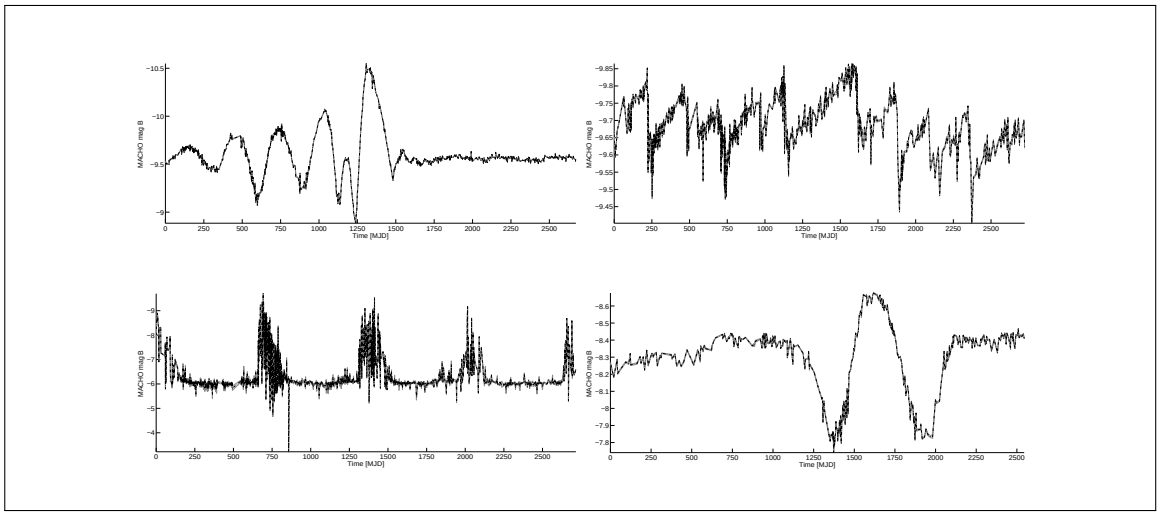

FIGURE 5.16. Top left panel Outlier MACHO_13.5835.11, top right panel Outlier MACHO_18.2478.9, bottom left panel Outlier MACHO_78.6462.561 and bottom right panel Outlier MACHO_62.7241.19.

6. CONCLUSIONS

The generation of precise, large and complete sky surveys in the last 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 anomaly among other applications. In this work we presented an algorithm based on a supervised classifier mechanism that enables us to discover outliers in catalogs of light-curves. Different from the existing methods, our work develops 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 overtake of the resources. Nevertheless, our algorithm is only expensive in the training stage and extremely fast analyzing the unknown light-curves, allowing us to explore a very large dataset. Furthermore, our method is not only restricted to astronomical problems, being applicable to any data base where anomaly detection is necessary.

The results from the application of our work on catalogs of classified periodic stars from MACHO project are encouraging, showing 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. The artifacts were removed from the outlier list and added to the training set. After retraining, we cross-matched the new candidates with the available catalogs and found known but rare objects among our outliers and also objects that did not have previous information. We

classified some of them as new variability classes and others as intriguing unique outliers. As a 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 will be able to find more of these objects and turn our isolated outliers into new known variability classes.

Furthermore, we are planning to improve our algorithm by constructing a more complete and large training set and by creating new robust features. Our future approach will mainly belong to the family of deep learning and unsupervised feature learning techniques, where the most representative patterns from objects are automatically discovered to represent every object as a mixture of these patterns. We also aim to apply our algorithm to different large sky surveys 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 a software which will include feature calculation of the light-curves and the application of our algorithm as a downloadable software and as an on-line tool and web services in the near future.

References

Agarwal, D. (2005). An empirical bayes approach to detect anomalies in dynamic multidimensional arrays. In *Proceedings of the 5th IEEE international conference on data mining. IEEE computer society* (p. 26-33).

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *ACM SIGMOD Record*, *30*(2), 37-46. Retrieved from http://portal.acm.org/citation.cfm?doid=376284.375668

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Becker, A. C., Bennett, D. P., ... Welch, D. (1997a). First Detection of a Gravitational Microlensing Candidate toward the Small Magellanic Cloud. *The Astrophysical Journal Letters*, *491*, L11.

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Becker, A. C., Bennett, D. P., ... MACHO Collaboration (1997b). The MACHO Project Large Magellanic Cloud Microlensing Results from the First Two Years and the Nature of the Galactic Dark Halo. *The Astrophysical Journal*, *486*, 697.

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Becker, A. C., Bennett, D. P., ... Welch, D. L. (1997c). The MACHO Project Large Magellanic Cloud Variable Star Inventory. III. Multimode RR Lyrae Stars, Distance to the Large Magellanic Cloud, and Age of the Oldest Stars. *The Astrophysical Journal*, *482*, 89.

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Becker, A. C., Bennett,

D. P., ... Welch, D. L. (1997d). The MACHO Project LMC Variable Star Inventory.V.Classification and Orbits of 611 Eclipsing Binary Stars. *The Astronomical Journal*, *114*, 326.

Alcock, C., Allsman, R. A., Alves, D., Axelrod, T. S., Bennett, D. P., Cook, K. H., ... Sutherland, W. (1997). The MACHO Project: 45 Candidate Microlensing Events from the First-Year Galactic Bulge Data. *The Astrophysical Journal*, *479*, 119.

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., & Bennett, D. P. (2001). Astrometry with the MACHO Data Archive. I. High Proper Motion Stars toward the Galactic Bulge and Magellanic Clouds. *The Astrophysical ...*, *20*(2000). Retrieved from http://iopscience.iop.org/0004-637X/562/1/337

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., ... Welch, D. L. (1999). The MACHO Project LMC Variable Star Inventory. VIII. The Recent Star Formation History of the Large Magellanic Cloud from the Cepheid Period Distribution. *The Astronomical Journal*, *117*, 920-926.

Alcock, C., Allsman, R. A., Alves, D. R., Becker, A. C., Bennett, D. P., Cook, K. H., ... Welch, D. L. (2002). The MACHO Project Large Magellanic Cloud Variable Star Inventory. XII. Three Cepheid Variables in Eclipsing Binaries. *The Astrophysical Journal*, *573*, 338-350.

Alcock, C., Allsman, R. A., Axelrod, T. S., Bennett, D. P., Cook, K. H., Freeman, K. C., ... Welch, D. L. (1996). The MACHO Project LMC Variable Star Inventory.II.LMC RR Lyrae Stars- Pulsational Characteristics and Indications of a Global Youth of the LMC. *The Astronomical Journal*, *111*, 1146.

Alcock, C., Alves, D. R., Axelrod, T. S., Becker, a. C., Bennett, D. P., Clement, C. M., ... Welch, D. L. (2004). The MACHO Project Large Magellanic Cloud Variable-Star Inventory. XIII. Fourier Parameters for the First-Overtone RR Lyrae Variables and the LMC Distance. *The Astronomical Journal*, *127*(1), 334-354. Retrieved from http://stacks.iop.org/1538-3881/127/i=1/a=334

Ansari, R. (2004). Eros: a galactic microlensing odyssey. In *International conference on cosmic rays and dark matter* (pp. 1–9).

Arning, A., Agrawal, R., & Raghavan, P. (1996). A linear method for deviation detection in large databases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 164-169).

Artyukhina, N., Durlevich, O., Frolov, M., Goranskij, V., Gorynya, N., Karitskaya, E., ... others (1996). Gcvs, vol. v.: Extragalactic variable stars (artyukhina+ 1996). *VizieR Online Data Catalog*, *2205*, 0.

Bhattacharyya, S., Richards, J. W., Rice, J., Starr, D. L., Butler, N. R., & Bloom, J. S. (2012). Identification of Outliers through Clustering and Semi-supervised Learning In All Sky Automated Survey Data Set. , 1-2.

Bishop, C. (1994). Novelty detection and neural network validation. In *Proceedings of iee conference on vision, image and signal processing.* (p. 217-222).

Blanco, V. M., & Heathcote, S. (1986). A Quasar in the Direction of the Large Magellanic Cloud. *PASP*, *98*, 635.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning*

theory (p. 144-152).

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data* (p. 93-104).

Cassisi, S., & Salaris, M. (2011). A Classical Cepheid in a Large Magellanic Cloud Eclipsing Binary: Evidence Of Shortcomings in Current Stellar Evolutionary Models? *The Astrophysical Journal Letters*, *728*, L43.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. *ACM Computing Surveys*, *41*(3), 1-58.

Chen, W. P., Sanchawala, K., & Chiu, M. C. (2006). W Ursae Majoris Contact Binary Variables as X-Ray Sources. *The Astronomical Journal*, *131*, 990-993.

Clayton, G. C. (1996). The R Coronae Borealis Stars. *Publications of the Astronomical Society of the Pacific*, 225-241.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309-347.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, *27*(3), 326-327.

Dobrzycki, A., Groot, P. J., Macri, L. M., & Stanek, K. Z. (2002). Discovery of Four X-Ray Quasars behind the Large Magellanic Cloud. *The Astrophysical Journal Letters*,

569, L15-L18.

Done, C., Gierliński, M., & Kubota, A. (2007). Modelling the behaviour of accretion flows in X-ray binaries. Everything you always wanted to know about accretion but were afraid to ask. *The Astronomy and Astrophysics Review*, *15*, 1-66.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning.* (p. 255-262).

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, p. 226-231).

Fender, R. P., Belloni, T. M., & Gallo, E. (2004). Towards a unified model for black hole X-ray binary jets. *Monthly Notices of the Royal Astronomical Society*, *355*, 1105-1118.

Fraser, O. J., Hawley, S. L., & Cook, K. H. (2008). the Properties of Long-Period Variables in the Large Magellanic Cloud From Macho. *The Astronomical Journal*, *136*(3), 1242-1258.

Geha, M., Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., ... Welch, D. L. (2003). Variability-selected Quasars in MACHO Project Magellanic Cloud Fields. *The Astronomical Journal*, *125*, 1-12.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3-42. Retrieved from http://link.springer.com/10.1007/s10994-006-6226-1

Gibbons, P. B., & Matias, Y. (1998). New sampling-based summary statistics for improving approximate query answers. In *ACM SIGMOD Record* (Vol. 27, p. 331-342).

Groenewegen, M. (2004). Long period variables in the magellanic clouds: Ogle+ 2 mass+

denis. *Astronomy and Astrophysics*, *425*, 595–613.

Grubb, & Frank, E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, 100-108.

He, J., & Carbonell, J. (2006). Rare Class Discovery Based on Active Learning.

Henrion, M., Hand, D. J., Gandy, A., & Mortlock, D. J. (2013). Casos: a subspace method for anomaly detection in high dimensional astronomical databases. *Statistical Analysis and Data Mining*, *6*(1), 53-72.

Herschel, J. (1857). *Outlines of astronomy*. Blanchard and Lea. Retrieved from http://books.google.com/books?hl=en\&lr=\&id=3LsMAAAAYAAJ\&oi=fnd\&pg=PA17\&dq=Outlines+of+astronomy\&ots=oa5ap5Uks7\&sig=LUlnosl8e6Q9mDWjbFjMp1HXsS8

Hodapp, K. W., Kaiser, N., Aussel, H., Burgett, W., Chambers, K. C., Chun, M., ... Waterson, M. (2004). Design of the Pan-STARRS telescopes. *Astronomische Nachrichten*, *325*(6-8), 636-642. Retrieved from http://doi.wiley.com/10.1002/asna.200410300

Huijse, P., Member, S., & Est, P. A. (2014). Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases. , 1-26.

Ita, Y., Tanabé, T., Matsunaga, N., Nakajima, Y., Nagashima, C., Nagayama, T., ...

Nakada, Y. (2004). Variable stars in the Magellanic Clouds - II. The data and infrared properties. *Monthly Notices of the Royal Astronomical Society*, *353*(3), 705-712. Retrieved from http://mnras.oxfordjournals.org/cgi/doi/10.1111/j.1365-2966.2004.08126.x

Jin, W., Tung, A., & Han, J. (2001). Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge discovery and data mining* (p. 293-298).

John, G. (1995). Robust decision trees: Removing outliers from databases. In *Proceedings of the first international conference on knowledge discovery and data mining* (p. 174-179).

Keller, S. C., Bessell, M. S., Cook, K. H., Geha, M., & Syphers, D. (2002). Blue Variable Stars from the MACHO Database. I. Photometry and Spectroscopy of the Large Magellanic Cloud Sample. *The Astronomical Journal*, *124*(4), 2039-2044. Retrieved from http://stacks.iop.org/1538-3881/124/i=4/a=2039

Keller, S. C., Schmidt, B. P., Bessell, M. S., Conroy, P. G., Francis, P., Granlund, A., ... others (2007). The skymapper telescope and the southern sky survey. *Publications of the Astronomical Society of Australia*, *24*(1), 1-12.

Keller, S. C., & Wood, P. R. (2002). Large Magellanic Cloud Bump Cepheids: Probing the Stellar Mass-Luminosity Relation. *The Astrophysical Journal*, *578*, 144-150.

Kim, D.-W., Protopapas, P., Bailer-Jones, C., Byun, Y.-I., Chang, J.-B., Marquette, J.-B., & Shin, M.-S. (2014). The EPOCH Project: I. Periodic Variable Stars in the EROS-2 LMC Database. *submitted to Astronomy and Astrophysics*.

Kim, D.-W., Protopapas, P., Byun, Y.-I., Alcock, C., Khardon, R., & Trichas, M. (2011). Quasi-Stellar Object Selection Algorithm Using Time Variability and Machine Learning: Selection of 1620 Quasi-Stellar Object Candidates From Macho Large Magellanic Cloud Database. *The Astrophysical Journal*, *735*(2), 68.

Kim, D.-W., Protopapas, P., Trichas, M., Rowan-Robinson, M., Khardon, R., Alcock, C., & Byun, Y.-I. (2012). a Refined Qso Selection Method Using Diagnostics Tests: 663 Qso Candidates in the Large Magellanic Cloud. *The Astrophysical Journal*, *747*(2), 107.

Knigge, C. (2011). The Evolution of Cataclysmic Variables. In L. Schmidtobreick, M. R. Schreiber, & C. Tappert (Eds.), *Evolution of compact binaries* (Vol. 447, p. 3).

Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the VLDB Conference, New York, USA* (p. 392-403).

Koller, D., & Friedman, N. I. R. (2009). *Probabilistic graphical models*.

Kou, Y., Lu, C., Sirwongwattana, S., & Huang, Y. (2004). Survey of fraud detection techniques. In *Proceedings of the IEEE international conference on networking, sensing and control* (p. 749-754).

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. (2005). High-mass X-ray binaries in the Magellanic Clouds. *Astronomy and Astrophysics*, *442*, 1135-1138.

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. (2007). A catalogue of low-mass X-ray binaries in the Galaxy, LMC, and SMC (Fourth edition). *Astronomy and Astrophysics*, *469*, 807-810.

Marconi, M., Molinaro, R., Bono, G., Pietrzyński, G., Gieren, W., Pilecki, B., ... Karczmarek, P. (2013). The Eclipsing Binary Cepheid OGLE-LMC-CEP-0227 in the Large

Magellanic Cloud: Pulsation Modeling of Light and Radial Velocity Curves. *The Astrophysical Journal Letters*, *768*, L6.

Monti, S., & Cooper, G. F. (1998). A multivariate discretization method for learning bayesian networks from mixed data. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (p. 404-413).

Nairac, A., Towsend, N., Carr, R., King, S., Cowley, P., & Tarassenko, L. (1999). A system for the analisys of jet system vibration data. *Integrated ComputerAided Engineering*, *6*(1), 53-65.

Ness, J.-U., Schmitt, J. H. M. M., Burwitz, V., Mewe, R., & Predehl, P. (2002). Chandra LETGS observation of the active binary Algol. *Astronomy and Astrophysics*, *387*, 1032-1046.

Nun, I., Pichara, K., Protopapas, P., & Kim, D.-W. (2014). Supervised detection of anomalous light-curves in massive astronomical catalogs. *arXiv preprint arXiv:1404.4888*, 17. Retrieved from http://arxiv.org/abs/1404.4888

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. In *Data engineering, 2003. proceedings. 19th international conference on* (p. 315-326).

Penzias, A., & Wilson, R. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 419-421.

Percy, J. R. (2007). *Understanding variable stars*. Cambridge University Press.

Pichara, K., & Protopapas, P. (2013). Automatic classification of variable stars in catalogs with missing data. *The Astrophysical Journal*, *777*(2), 83.

Pichara, K., Protopapas, P., Kim, D.-W., Marquette, J.-B., & Tisserand, P. (2012). An improved quasar detection method in eros-2 and macho lmc data sets. *Monthly Notices of the Royal Astronomical Society*, *427*(2), 1284-1297.

Pichara, K., & Soto, A. (2011). Active learning and subspace clustering for anomaly detection. *Intelligent Data Analysis*, *15*(2), 151-171.

Pichara, K., Soto, A., & Araneda, A. (2008). Detection of anomalies in large datasets using an active learning scheme based on dirichlet distributions. In *Advances in artificial intelligence-iberamia 2008* (p. 163-172). Springer.

Pietrzyński, G., Thompson, I. B., Gieren, W., Graczyk, D., Bono, G., Udalski, A., ... Pilecki, B. (2010). The dynamical mass of a classical Cepheid variable star in an eclipsing binary system. *Nature*, *468*, 542-544.

Poleski, R. (2008). Period changes of lmc cepheids in the ogle and macho data. *Acta Astronomica*, *58*, 313.

Protopapas, P., Giammarco, J., Faccioli, L., Struble, M., Dave, R., & Alcock, C. (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, *369*(2), 677-696.

Quimby, R. M., Kulkarni, S. R., Kasliwal, M. M., Gal-Yam, A., Arcavi, I., Sullivan, M., ... Levitan, D. (2011). Hydrogen-poor superluminous stellar explosions. *Nature*, *474*, 487-489.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD Conference on Management*

of Data, Dallas, TX (p. 427-438).

Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. (2008). Finding anomalous periodic time series. *Machine Learning*, *74*(3), 281-313. Retrieved from http://link .springer.com/10.1007/s10994-008-5093-3

Reynolds, D. (2009). Gaussian mixture models. *Encyclopedia of Biometrics*, 659-663.

Richards, J. W., Starr, D. L., Miller, A. A., Bloom, J. S., Butler, N. R., Brink, H., & Crellin-Quick, A. (2012). Construction of a calibrated probabilistic classification catalog: Application to 50k variable sources in the all-sky automated survey. *The Astrophysical Journal Supplement Series*, *203*(2), 32.

Ridley, J. P., Crawford, F., Lorimer, D. R., Bailey, S. R., Madden, J. H., Anella, R., & Chennamangalam, J. (2013). Eight new radio pulsars in the Large Magellanic Cloud. *Monthly Notices of the Royal Astronomical Society*, *433*, 138-146.

Ruiz, M., Mujica, L. E., Berjaga, X., & Rodellar, J. (2013). Partial least square/projection to latent structures (pls) regression to estimate impact localization in structures. *Smart Materials and Structures*, *22*(2), 025028. Retrieved from http://stacks.iop.org/0964-1726/22/i=2/a=025028

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). *Learning representations by back-propagating errors*. MIT Press, Cambridge, MA, USA.

Schaefer, B. E. (2010). Comprehensive Photometric Histories of All Known Galactic Recurrent Novae. *The Astrophysical Journal Supplement Series*, *187*, 275-373.

Schaubert, D., Boryssenko, A., Van Ardenne, A., de Vaate, J. B., & Craeye, C. (2003). The square kilometer array (ska) antenna. In *Phased array systems and technology, 2003.*

IEEE international symposium on (p. 351-358).

Schmidtke, P. C., Cowley, A. P., Crane, J. D., Taylor, V. A., McGrath, T. K., Hutchings, J. B., & Crampton, D. (1999). Magellanic Cloud X-Ray Sources. III. Completion of a ROSAT Survey. *The Astronomical Journal*, *117*, 927-936.

Seidl, T., Müller, E., Assent, I., & Steinhausen, U. (2009). Outlier detection and ranking based on subspace clustering. In *Uncertainty management in information systems.*

Serio, G. F., Manara, A., Sicoli, P., & Bottke, W. F. (2002). *Giuseppe piazzi and the discovery of ceres*. University of Arizona Press.

Shafter, A. W. (2013). Photometric and Spectroscopic Properties of Novae in the Large Magellanic Cloud. *The Astronomical Journal*, *145*, 117.

Son, C., Cho, S., & Yoo, J. (2009). Volume traffic anomaly detection using hierarchical clustering. In *Management enabling the future internet for changing business and new computing services* (Vol. 5787, p. 291-300). Springer Berlin, Heidelberg.

Soszynski, I., Udalski, A., Kubiak, M., Zebrun, K., & Szewczyk, O. (2003). The optical gravitational lensing experiment. catalog of rr lyr stars in the large magellanic cloud£. *Acta Astronomica*, *53*, 93–116.

Soszynski, I., Udalski, A., Szymanski, M., Kubiak, M., Pietrzynski, G., Wyrzykowski, O., ... Poleski, R. (2008). The optical gravitational lensing experiment. the ogle-iii catalog of variable stars. ii. type ii cepheids and anomalous cepheids in the large magellanic cloud. *Acta Astronomica*, *58*, 293–312.

Thomas, C. L., Griest, K., Popowski, P., Cook, K. H., Drake, A. J., Minniti, D., ... MACHO Collaboration (2005). Galactic Bulge Microlensing Events from the MACHO

Collaboration. *The Astrophysical Journal*, *631*, 906-934.

Tyson, J. A., Collaboration, L., Labs, B., Technologies, L., & Hill, M. (2002). Large Synoptic Survey Telescope : Overview. *Astronomical Telescopes and Instrumentation*, 10-20.

Udalski, A., Kubiak, M., & Szymanski, M. (1997). Optical gravitational lensing experiment. ogle-2–the second phase of the ogle project. *Acta Astronomica*, *47*, 319–344.

van den Heuvel, E. P. J., Bhattacharya, D., Nomoto, K., & Rappaport, S. A. (1992). Accreting white dwarf models for CAL 83, CAL 87 and other ultrasoft X-ray sources in the LMC. *Astronomy and Astrophysics*, *262*, 97-105.

van der Klis, M. (2000). Millisecond Oscillations in X-ray Binaries. *Astronomy and Astrophysics*, *38*, 717-760.

Voges, W., Aschenbach, B., Boller, T., Bräuninger, H., Briel, U., Burkert, W., ... others (1999a). The rosat all-sky survey bright source catalogue (1rxs). *Astronomy and Astrophysics*, *349*, 389–405.

Voges, W., Aschenbach, B., Boller, T., Bräuninger, H., Briel, U., Burkert, W., ... others (1999b). The rosat all-sky survey bright source catalogue (1rxs). *Astronomy and Astrophysics*, *349*, 389–405.

Wall, J. V., & Jenkins, C. R. (2012). *Practical statistics for astronomers*. Cambridge University Press.

Watson, M., Schröder, A., Fyfe, D., Page, C., Mateos, S., Pye, J., ... others (2009). The xmm-newton serendipitous survey: V. the second xmm-newton serendipitous source catalogue. *Astronomy and Astrophysics*, *493*(1), 339–373.

Watson, M. G., Schröder, A. C., Fyfe, D., Page, C. G., Lamer, G., Mateos, S., ... Yuan, W. (2009). The XMM-Newton serendipitous survey. V. The Second XMM-Newton serendipitous source catalogue. *Astronomy and Astrophysics*, *493*, 339-373.

Wood, P. R. (2000). Variable red giants in the LMC: Pulsating stars and binaries? *PASP*, *17*, 18-21.

Wyrzykowski, L., Udalski, A., Kubiak, M., Szymanski, M., Zebrun, K., Soszynski, I., ... Szewczyk, O. (2003). The optical gravitational lensing experiment. eclipsing binary stars in the large magellanic cloud. *Acta Astronomica*, *53*, 1–25.

Xiong, L., Poczos, B., Connolly, A., & Schneider, J. (2010). *Anomaly detection for astronomical data.* December.

Yang, H., Xie, F., & Lu, Y. (2006). Clustering and classification based anomaly detection. In *Fuzzy systems and knowledge discovery* (Vol. 4223, p. 1082-1091).

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., ... Berman, E. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, *120*(3), 1579.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (p. 103-114).