UMBC

AN HONORS UNIVERSITY IN MARYLAND

JCA

Predictive Mining of Time Series Data

Akshay Java (Computer Science Department)
Eric Perlman (Joint Center for Astrophysics, Physics Department)

ABSTRACT

All-sky monitors are a relatively new development in astronomy, and their data represent a largely untapped resource. Proper utilization of this resource could lead to important discoveries not only in the physics of variable objects, but in how one observes such objects. We discuss the development of a Java toolbox for astronomical time series data. Rather than using methods conventional in astronomy (e.g., power spectrum and cross-correlation analysis) we employ rule discovery techniques commonly used in analyzing stock-market data. By clustering patterns found within the data, rule discovery allows one to build predictive models, allowing one to forecast when a given event might occur or whether the occurrence of one event will trigger a second. We have tested the toolbox and accompanying display tool on datasets (representing several classes of objects) from the RXTE All Sky Monitor. We use these datasets to illustrate the methods and functionality of the toolbox. We have found predictive patterns in several ASM datasets. We also discuss problems faced in the development process, particularly the difficulties of dealing with discretized and irregularly sampled data. A possible application would be in scheduling target of opportunity observations where the astronomer wants to observe an object when a certain event or series of events occurs. By combining such a toolbox with an automatic, Java query tool which regularly gathers data on objects of interest, the astronomer or telescope operator could use the real-time datastream to efficiently predict the occurrence of (for example) a flare or other event. By combining the toolbox with various preprocessing and dimensionality reduction tools, one could predict events which may happen on variable time scales.

1. Introduction

Many types of variable objects exist in the universe, including stars with predictable behavior (e.g., Cepheids), objects with behavior that is inherently unpredictable (e.g, AGN), and objects with both predictable and irregular variability patterns (e.g., X-ray binaries). Constant monitoring of variable objects has been a continuing interest in astronomy, beginning with 16th century astronomer David Fabricius, and extending through history to Herschel, Leavitt and others. Today, monitoring is done by a wide variety of techniques, observers and instruments, from dedicated amateurs, to professional astronomers interested in intensive monitoring of individual objects, to all-sky monitors such as te *RXTE ASM* and *BATSE* aboard *CGRO*.

All-sky monitoring projects are new, relatively untapped tools. Already they have made important, if not decisive contributions to solving some of astronomy's most persistent mysteries, such as the cosmological origin of gamma-ray bursts and linking different emission regions in AGN. All-sky monitors are an important resource, for studying the physics behind variability and also as tools to optimize the scheduling of observing programs designed either to study the behavior of objects in their active stages or to take advantage of the increased brightness of an object for particularly deep observations. With major initiatives such as the *Large-Area Synoptic Survey Telescope (LSST)* and *Supernova Acceleration Probe (SNAP)*, all-sky monitors are also poised to become a major discovery tool in astronomy, not only finding large numbers of variable objects of known classes, but perhaps also finding new classes of variable objects.

In order to maximize the utility of all-sky monitors, it is important to devise ways of handling large amounts of data in real time and find not variability and predictive patterns among these large data streams. Such tools could optimize the scheduling of programs designed to observe an object in a particular state, aid in the understanding of the lightcurves of known variable objects and facilitate the discovery of patterns in the lightcurves of new classes of variable objects. It was with these goals in mind that we undertook this project.

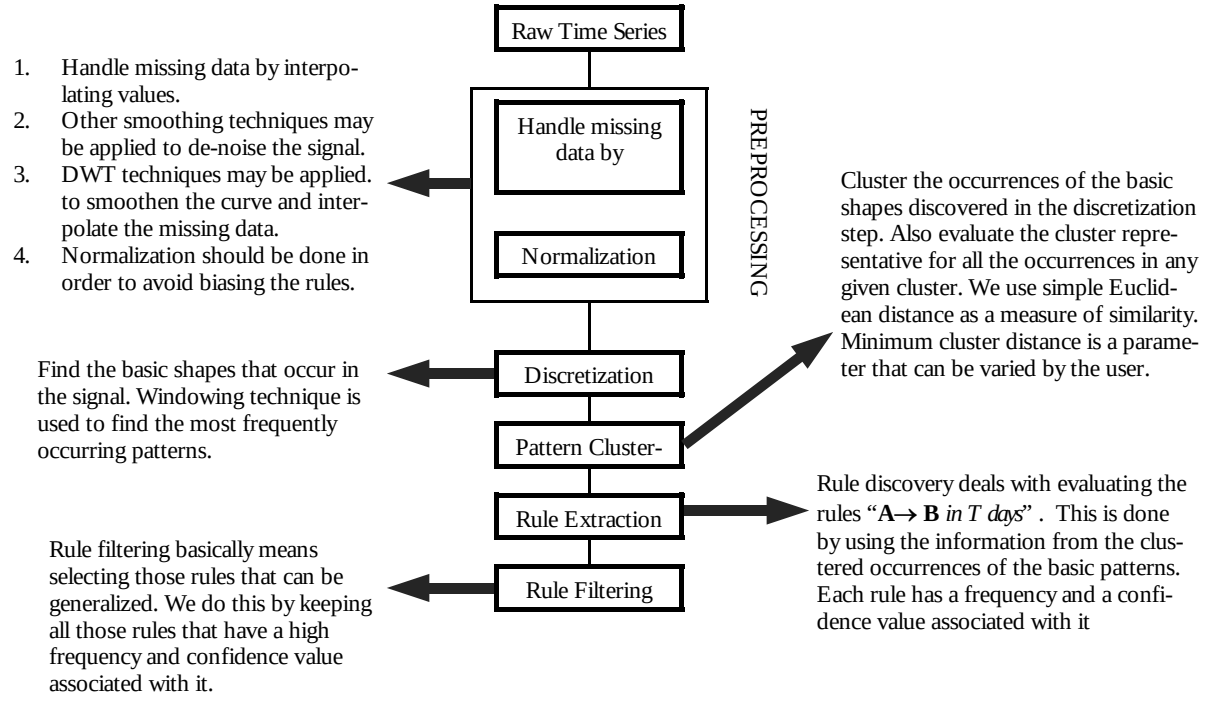

Figure 1. A flowchart that describes the basic Rule discovery Algorithm

2. Rule Discovery

The traditional problem in understanding variable objects is to find patterns within their lightcurves (i.e., time series). The need to understanding time series data is not unique to astronomy: time series data exists in a wide variety of fields including geology and atmospheric science as well as many business applications (e.g., stock market forecasting). The common problem in all of these diverse fields of inquiry is how to efficiently find patterns in the data.

Traditional analysis methods in astronomy have included Fourier transform and power spectrum methods which are well suited for finding patterns with relatively well-determined periodicities (for a review, see Scargle 1997). However those techniques may be less helpful for understanding variable objects with irregular behavior, and are not optimized to serve as predictive tools. Our approach to building a tool for analyzing time series data was to employ data-mining techniques which are optimized for finding patterns that do not rely on a regular periodicity. These include clustering and rule discovery, tools that are already in use in business applications and which are active, ongoing research topics in computer science.

Clustering and rule discovery are distinct data-mining techniques which are often complementary. The goal of clustering is to find patterns in any data set, while the goal of rule discovery is to find predictive patterns. This section is devoted to explaining the techniques and algorithms we employ.

2.1 Time series discretization by clustering

Our first goal is to find a representation of the dataset as a collection of patterns. One can understand the goal as being similar to that of Taylor series or eigenvector representation. We achieve this by first sliding a window of size w through the existing time series to get subsequences of size w each of which can be considered as a point in w-dimensional space. If there are n points in the time series the number of subsequences obtained after windowing is $n-w+1$.

These points are then clustered using any of the standard clustering algorithms (for the project we have implemented the greedy clustering algorithm explained in Das, Gunopolous & Mannila 1997). Once a good fit is achieved, these clusters can then be considered as the basic shapes of the time series, with the entire series composed of superpositions and/ or combinations of these basic shapes.

2.2 Rule discovery

The next and main step of the process is trying to find interesting probabilistic rules between the clusters in the two time series. These rules are of the form: "If a cluster A occurs in time series 1 then we can say with confidence c that B will occur in time series 2 within time T". This can be represented as
$(A \rightarrow B)(T)$, and the confidence of this rule is given by

$$c((A \rightarrow B)(T)) = F((A \rightarrow B)(T))/F(A)$$

where $F(A)$ is the frequency of cluster A in time series 1 and $F((A \rightarrow B)(T))$ is the number of occurrences of cluster A in time series 1 which are followed by an occurrence of cluster B in time series 2 in time at most T.

The entire technique is based on the selection of a number of parameters such as the window size *w*, the parameter used in clustering, the time *T* in the above rule etc., the optimum values of which can be obtained only after empirical testing.

Figure 1 displays the basic algorithm used in predictive time series analysis. This is an implementation of the technique described by Das *et al.* (1996, 1997). The raw time series is first processed by handling the missing data by interpolation and normalizing the time series. Normalization is an important step that is used to avoid any biasing in the rules.

The results of the rule discovery program are rules of the form $A \rightarrow B$ in T days with frequency f and c as confidence. This may be interpreted as "an occurrence of pattern A is followed by the occurrence of pattern B in T days". The frequency is the total number of occurrences of such rules in the time series. The confidence can be defined as the probability of occurrence of the 2nd pattern within T days given that the first has occurred. Ideally, one wants rules with both high frequency and a high confidence. This can be combined with the signal to noise ratio or error bars to filter the results.

2.3 Extensions for Multivariate time series Analysis

Most data in the real world are interdependent on various factors. *RXTE ASM* datasets, are multivariate time series with attributes such as sum band intensity, A band intensity etc. In order to observe the influence of one feature on the other we can modify the above algorithm such that the first shape A is selected from one attribute and the second shape B is selected from the other time series. So this would show the occurrence of the pattern A in one time series being followed by a pattern B in the other time series. The implication of this is that the scheme allows us to select the relevant attributes or performing a feature selection to study those features that could strongly influence (for example) the intensity of the object.

2.4 Filtering of rules

The basic scheme for filtering the rules is by looking at the frequency and the confidence values associated with each of these rules. The rules that have a high frequency and high confidence value are the ones that convey information about the most general rules that apply to the entire time series. However one might be also interested in identifying rare events that might occur with low frequency but with a high confidence value. These rules might be the most useful in scheduling the target of opportunity events for rare occurrences that might otherwise go unnoticed. Rules may also be filtered using the signal to noise ratio. Rules that have a high value of signal to noise ratio are the ones that might convey more information since they are less influenced by errors. Another variation to the filtering technique could be to use relevance feedback mechanisms by which the user may assign certain weights to the patterns that he/ she might consider as most interesting for his observations. Another form of relevance feedback mechanism is by using the quality of the predictive nature of the rules as a feedback for the inference algorithm. Basic shapes associated with rules that are more predictive in nature are associated with higher weight values.

2.5 Other issues

Another interesting issue in the predictive rule discovery technique is the extension of this basic algorithm to rule discovery for sequence of occurrences of a pattern. Thus what we may infer from this would be rules of the form " *n* occurrences of the basic pattern of shape A are followed by *m* occurrences of the basic pattern of shape B in the next T days". This may be particularly useful in the cases where the time series have a large fluctuations and we need to mine for predictive patterns (see Agarwal 1993).

3. Data Analysis Structure and Usage

Figure 2 shows the overall structure of data analysis using this program. Two goals are possible, each requiring a slightly different analysis tree: (1) to schedule a telescope more intelligently to objects when they are at their brightest or most interesting stage, or (2) as an analysis tool to find patterns within a (possibly multivariate) time series. Both branches are shown in the diagram, although it should be noted that they are largely similar with the exception of interaction with the scheduling process of a given telescope. In the discussion below, we will concentrate on the use of the algorithm for scheduling a telescope, as from Figure 2 it is relatively straightforward to envision the application to an individual investigator.

The Time Series Analysis (TSA) Module is the heart of the system where the actual data are monitored in real-time. Within the TSA, the user can control three main parameters: the window size, the minimum cluster distance and the time period for prediction. The system comprises of two parts each of which works in conjunction with the planner and the scheduler in deciding which object is to be observed. The TSA module analysis the time series data and performs trend prediction and rule discovery on both real time and historical data. The scheduler uses this information along with information on object observability to forecast not only the activity level of an object but also the ability of the telescope or satellite to perform the desired observations. All of these factors would be considered in the light of the constraints imposed by individual investigators as well as the review committee.

Hence for example when we have two programs where one object is at a brightness of 60% [compared to some historical maximum] and the other object is at 90% brightness, we may be able schedule the telescope more intelligently if we were to know (for example) that based on the rules discovered within their time series, object B will be at 90% brightness again in the next 10 days while object A may not be at even 30% brightness in the next 2 years. However if we were to schedule the telescope only based on certain threshold values rather than rules discovered using time series data mining we would miss out on this information. Numerous examples of this having occurred exist in the astronomical literature. Another interesting example can be the ability of the system to be able to perform observations based on a particular trend in real time basis. Thus a rising or a falling trend that may indicate some interesting astronomical phenomenon that may otherwise have gone unnoticed can now be observed due to rule discovery.

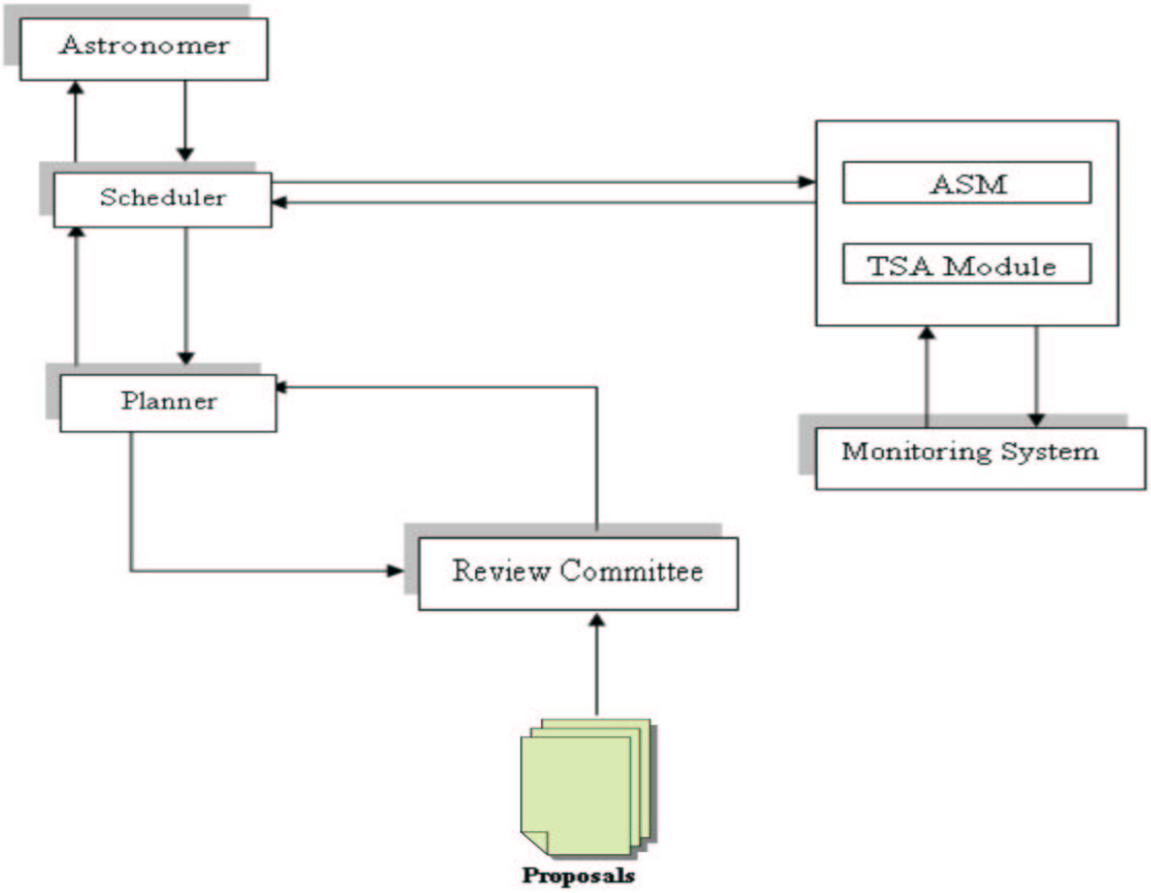

Figure 2. The usage and the analysis structure for incorporation of a real-time data analysis system within a satellite's scheduling process. This example shows an application in which real time datastreams may be monitored for observing interesting behavior. At the same time this tool would be able to help the scientist schedule and plan observations based on the behavioral patterns of the object that were found by the rule discovery technique.

4. Experimental Results

To test and validate the toolbox, we used RXTE ASM datasets for various objects, spanning several different object classes. Below and at right, we show one particularly interesting dataset, from SMC X-1, a bright X-ray binary in the Small Magellanic Cloud. As shown in Figure 3, SMC X-1 showed interesting periodic behavior on several different timescales: a long time-period behavior as well as short time-scale variations superposed upon the long-period behavior, but occurring only when the object is bright. This combination, which has been discussed in many papers (e.g., Kahabka & Li 1999) made it an ideal choice for the experimental setup. Through the experiments we observed that the toolbox was able to provide predictive patterns as shown in Figure 4.

The toolkit was found to be particularly effective for predicting rules over a short time span. The effect of varying window size was studied and the results are as shown in Figure 6 (left panel).

From the results it was observed that as we vary the window size we keep capturing all the basic patterns as we go along incrementing the window size. We observed that after a particular point increasing the window size does not help improve the number of useful results that we can obtain. In order to be able to also capture the rules that occur on a long term basis we need to use some smoothing techniques in order to be able to effectively find all the basic shapes occurring in the time series. Fourier techniques or other smoothing methods can be used for this purpose. For example by using Moving averages we were able to smoothen out the noise in the signal. This smoothened signal is then used for finding the rules and this can be treated as a preprocessing technique. Other techniques include Wavelet Transformation etc.

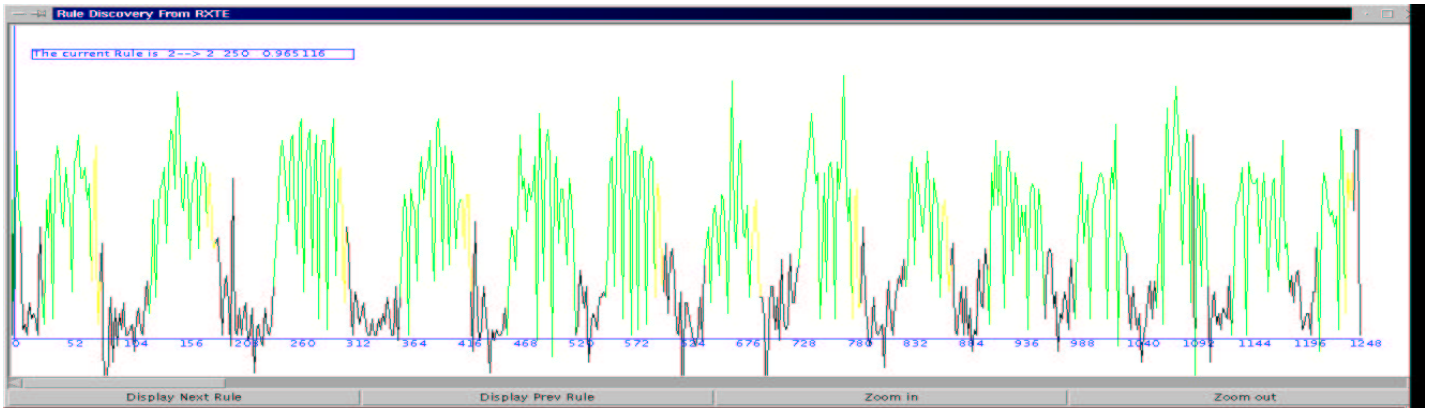

Figure 3. A screenshot of the Java GUI application that is used to display the rules generated. Each color represents an occurrence of a basic pattern. The rule represented by the above figure is of the form 2→2 which means that every occurrence of basic pattern 2 is followed by an occurrence of the same pattern within the next 20 days. It also shows the frequency with which the rule occurs and the confidence values associated with the rule.

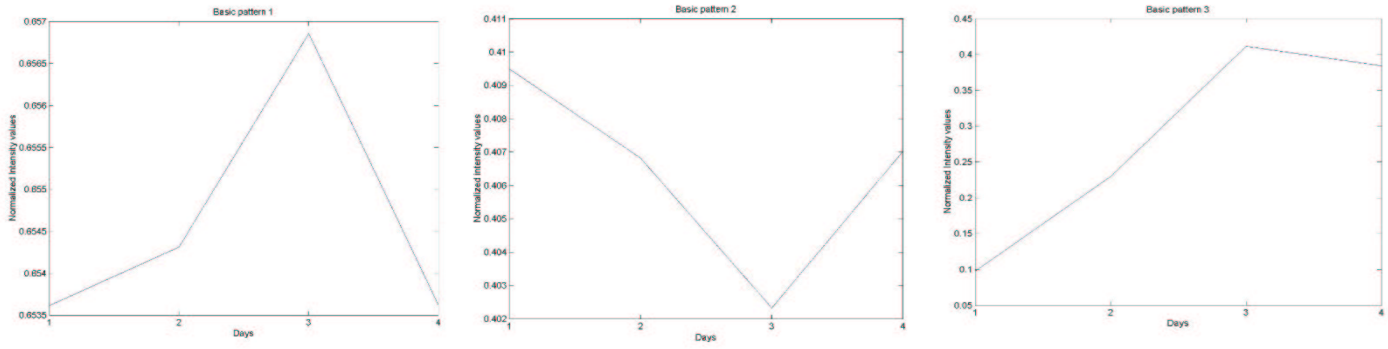

Figure 4. The basic shapes or patterns that were found by the algorithm in the SMC X-1 dataset. The occurrences of these basic shapes are then clustered and the corresponding rules are found. The above shapes for example were discovered with a window size of 4 days and the minimum cluster distance of 0.7. The shapes display the variations of the intensity of the object over a period of 4 days.

5. Challenges

5.1 Preprocessing: Dealing with Missing Data

In many cases there is a strong need to handle missing data effectively without damage to the results. For example in *RXTE ASM* datasets, points can be missing for several reasons including object observability as well as instrument performance.

There are several possible ways of dealing with missing data (see Agarwal et al. 1995 for a review of techniques). One is to simply interpolate between the values at each end of the missing data period. If the missing time period is small compared to the window size, this does not affect the rules obtained drastically. If, however, the missing time period is comparable to the window size, and if the number of data gaps is large, the interpolated periods can find their way into the list of significant rules. Such aliasing is obviously not desirable.

A second way of dealing with missing data is to use moving averages. This allows the use of local information to fill in gaps with approximately correct slope information and reduces the likelihood of aliasing. The figure below shows some results from this procedure. The moving averages technique can easily eliminate the effects of sudden noise fluctuations and at the same time approximate the curve for any missing values.

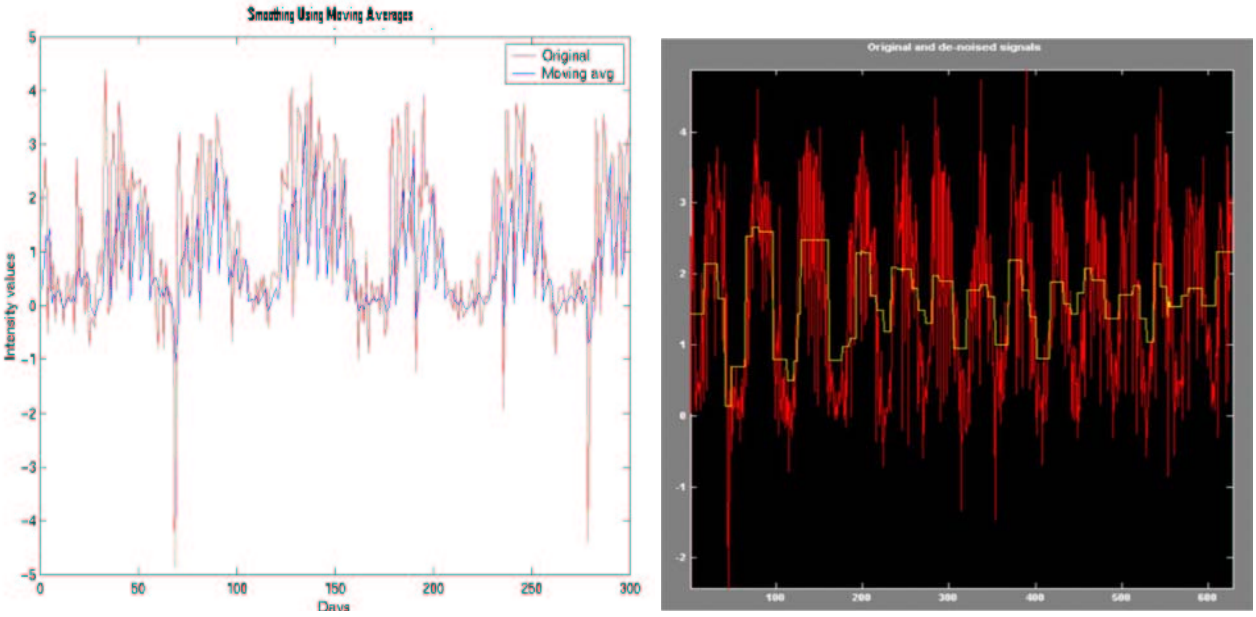

Figure 5. At left, we show preprocessing of the curve using moving averages. This technique can be used as a step to eliminate any sudden noise fluctuations and to compensate for any missing data points. At right, we show a wavelet approximation of the curve. The approximated curve captures the overall trend and can be used to discover rules that may explain the general behavior, or allow the separation of behaviors on multiple timescales.

5.2 Variability on multiple time-scales

An astronomical object can exhibit seasonal behavior on several different timescales. For example, in an X-ray binaries, there can be multiple periodicities connected both with the binary orbit as well as dynamical effects within the accretion disk around the compact massive object. As seen in Figure 3, SMC X-1 shows periodicity both on timescales of days and several weeks.

There are several ways of dealing with this issue. The first and most obvious way is to attempt to change the window size, i.e., the time period over which data points are examined for periodic behavior. We have experimented with doing so; our results appear mixed. As shown below, the number of significant rules appears to change for SMC X-1 with window size, with the peak number of significant rules appearing at window sizes of about five days. The relative character of the windows also changes with the window size. One can envision looping the routine, varying this parameter, and in the process maximizing the number and significance of the rules found.

Another possible method is to use dimensionality reduction techniques such as Discrete Fourier Transform (DFT) or Discrete Wavelet Transforms (DWT, Polikar 1996) in the preprocessing stages. The DFT is the easier of these to picture, as it breaks down any function into a simple series of sine/ cosine components. The DFT would find the longer time-period variability as a distinct peak. Then by removing that peak and performing the inverse transform we would renormalize the data and remove one of the two variability timescales, categorizing the period in the process. We have utilized both DFT and DWT techniques to find long-period seasonality in the SMC X-1 dataset and remove it. Figure 5 (right panel) shows the result of DWT analysis for the SMC X-1 lightcurve. As can be seen, the long time-scale variability is modeled relatively well, allowing the study of shorter time-scale variability.

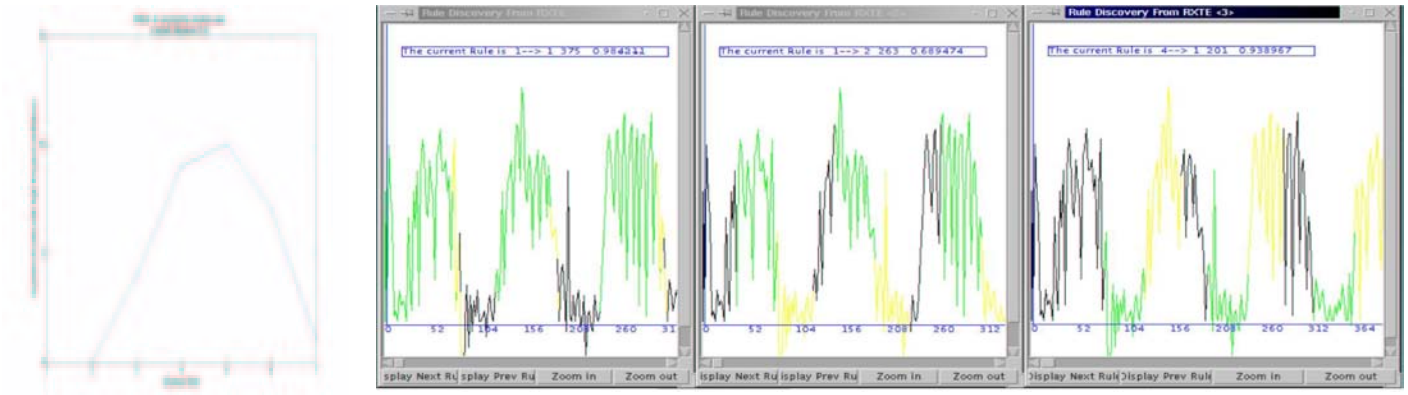

Figure 6. The leftmost panel shows the effect of varying the window size. As the size of the window increases the number of useful rules that are found may change. In order to be able to find rules on varying time scales we may need to de-noise or at times approximate the curve to get the overall trend information in it. The panels at right illustrate the changing character of the rules found as the window size changes from three (second from left) to five (second from right) to seven days (right).

5.3 Dimensionality reduction techniques for large time series databases

Typical application of time series data mining involves extremely large datasets and require certain techniques to cost effectively retrieve useful information from the data. A time series can be considered as a point in an *n*-dimensional data space. Dimensionality reduction techniques mathematically transform the entire time series into a lower dimensional representation in which the essence of the time series is preserved. By applying dimensionality reduction techniques such as singular value decomposition (SVD), DFT, DWT, and ARIMA (Auto Regressive Integrated Moving Average) we can efficiently find rules. (see Keogh et al. 2001 for a review on these techniques)

Dimensionality reduction is useful both at the preprocessing stage and also in the data analysis stage. When used as a preprocessing technique, it can allow the removal of unwanted (but known) periodicities from a datastream in order to facilitate the study of certain other features the user wants to study. When used as a data analysis tool, it can reduce a time series to an Eigenvector representation, perhaps as a prelude to Principal Components Analysis.

The other importance of such techniques is in feature selection in case of large multivariate time series. As already mentioned, many astronomical time series can contain interesting, periodic structure on a large variety of time scales. Often times one wants to pre-select features either for schedulability or science reasons. For example, in SMC X-1 the kilohertz QPOs are interesting scientifically but cannot at present be used for scheduling a telescope, while the structure on timescales of days to months can be used to schedule observations.

6. Future Work

The current implementation of the toolkit consist of a rule inference engine, programmed in C and a visualizer, programmed in Java. The present system can be used on any time-series dataset. However, the current toolkit has several limitations.

First, the current toolkit can analyze only a single time series at any time. This obviously gives it limited functionality, as many objects are studied with multiple instruments (indeed often times single instruments can yield variability data in several different bands, as can the *RXTE ASM*). Future implementations should include a multivariate time series analysis module, both for correlation as well as predictive purposes. This would enable the use of data from one source to predict interesting variability events in a completely different band, as is the goal of reverberation mapping of AGN (Horne et al. 2002).

A second significant improvement would be to incorporate dynamic time warping (DTW) modules (Keogh & Pazzani 2001). The utility of DTW is in spotting similar rules which may occur over a variety of timescales. This can be useful, for example if the object has quasi-periodic oscillations (QPOs), as have been seen in numerous X-ray binaries (van der Klis 1997) at kilohertz frequencies, too fast to affect scheduling. However, if QPOs related to accretion disk modes exist in AGN, the dynamical timescales would range from hours to weeks (Ulrich, Maraschi & Urry 1997) — therefore potentially having a major impact on scheduling. Unfortunately the record regarding QPOs in AGN has been mixed at best (see Benlloch et al. 2001 for a cautionary example).

A third improvement would be to incorporate various preprocessing or data analysis options (see Section 5 for examples) within the rule inference engine. This would significantly streamline the analysis of complex time-series data.

Finally, our eventual goal is to incorporate this into a scheduling system (as envisioned in Figure 2), enabling it to make use of incoming datastreams from all sky monitors and apply predictive rules to decide which observation should be scheduled next for satellites where pointed observations are required.

References

Agarwal, R., et al.., 1995, *Fast Similarity search in presence of noise scaling and translation in time series databases* in *Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland*

Agarwal, R., Inmielinski, T., Swami, A., 1993, *Mining Association Rules Between Sets of Items in Large Databases* in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*

Benlloch, S., et al., 2001, *Quasi-periodic Oscillation in Seyfert Galaxies: Significance Levels. The case of Markarian 766. Astrophysical Journal* 562, L121.

Das, G., Lin, K.-I., Mannila, H., Renganathan, G., Smyth, P., 1996, *Rule Discovery from Time Series* in *Proc. Of the 4th International Conference on Knowledge Discovery and Data Mining*, p. 16-22

Das, G., Gunopolous, D., Mannila, H., 1997 *Finding Similar Time Series* in *Principles of Data Mining and Knowledge Discovery*, p. 88-100

Hetland, M. L., 2000 *A survey of recent methods for efficient retrieval of similar time series*, Norwegian University of Science and Technology

Gunopulos, D., Das, G., 2000 *Time Series Similarity Measures* in *The Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*

Horne, K., et al., *Observational Requirements for High-Fidelity Reverberation Mapping*, submitted to *Publications of the Astronomical society of the Pacific*, astro-ph/ 0201182.

Kahabka, P., & Li, X-D, 1999, *The Recent Pulse Period Evolution of SMC X-1*, in *Astronomy & Astrophysics* 345, 117.

Keogh, E. J., et al. , 2001 *Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases* in *Knowledge and Information Systems* p. 263-286

Keogh, E. J., Pazzani, M. J., 2001, *Derivative Dynamic Time Warping* in *The First SIAM international conference on Data Mining*

van der Klis, M., 1997, *Kilohertz Quasi-Periodic Oscillations in Low Mass X-ray Binaries*, in *Astronomical Time Series*, ed. D. Maoz, A. Steinberg & E. M. Leibowitz (Kluwer), p. 121

Polikar, R., 1996, *The Wavelet Tutorial, second edition* http:/ / www.public.iastate.edu/ ~ rpolikar/ WAVELETS/ WTpart1.html

Scargle, J. D., 1997, *Astronomical Time Series Analysis: New Methods for Studying Periodic and Aperiodic Systems*. In *Astronomical Time Series*, ed. D. Maoz, A. Steinberg & E. M. Leibowitz (Dordrect: Kluwer), p. 1.

Ulrich, M.-H, Maraschi, L., Urry, C. M., 1997, *Variability of Active Galactic Nuclei*, in *Annual Review of Astronomy & Astrophysics*, 35, 445.