Machine-assisted discovery of relationships in astronomy

Matthew J. Graham,† S. G. Djorgovski, Ashish A. Mahabal, Ciro Donalek and Andrew J. Drake

California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, USA

ABSTRACT

High-volume feature-rich data sets are becoming the bread-and-butter of 21st century astronomy but present significant challenges to scientific discovery. In particular, identifying scientifically significant relationships between sets of parameters is non-trivial. Similar problems in biological and geosciences have led to the development of systems which can explore large parameter spaces and identify potentially interesting sets of associations. In this paper, we describe the application of automated discovery systems of relationships to astronomical data sets, focusing on an evolutionary programming technique and an information-theory technique. We demonstrate their use with classical astronomical relationships – the Hertzsprung–Russell diagram and the Fundamental Plane of elliptical galaxies. We also show how they work with the issue of binary classification which is relevant to the next generation of large synoptic sky surveys, such as the Large Synoptic Survey Telescope (LSST). We find that comparable results to more familiar techniques, such as decision trees, are achievable. Finally, we consider the reality of the relationships discovered and how this can be used for feature selection and extraction.

Key words: methods: data analysis – astronomical data bases: miscellaneous – virtual observatory tools.

1 INTRODUCTION

The rate of scientific discovery in astronomy has traditionally been tied to the amount of data available. The advent of digital astronomy with modern detectors and computational resources, e.g. data bases, has changed this. Although more data in the past two decades has allowed us to discover the cosmic web, dark energy and exoplanets, the vast majority of such low-hanging fruit have now been found. The new challenge is growing data complexity. The era of data-intensive astronomy promises a vastly more thorough exploration of parameter space but the discovery of new scientifically significant relationships is equally more complicated and daunting when faced with overwhelming data dimensions and volumes.

Given a highly complex data set, such as the Sloan Digital Sky Survey (SDSS), with hundreds of parameters for each object, and sufficient numbers of objects – hundreds of millions or more – to provide a fair and representative covering of parameter space, one will uncover many significant relationships – linear, non-linear, functional, structural – between pairs, triplets and groups of parameters. However, only a small fraction of these will be truly causative, the result of some valid underlying astrophysical process or processes, and identifying these is non-trivial. In fact, Cubitt, Eiser & Wolf (2012) have shown that identifying the underlying dynamical equations from any amount of experimental data, however precise, is a provably computationally hard (NP-hard1) problem.

The framework of astroinformatics, combining astronomy, applied computer science and information technology, has arisen to contend with this computational intractability. At its core are sophisticated data mining and multivariate statistical techniques which seek to extract and refine information from highly complex data sets (see Ball & Brunner 2010 for an overall review of different techniques in astronomy, Bloom & Richards 2012 for those specific to the time domain and the IVOA KDDIG webpages2 for general material related to this). This includes identifying unique or unusual classes of objects, estimating correlations and computing the statistical significance of a fit to a model in the presence of missing or bounded data, i.e. with lower or upper limits, as well as visualizing this information in a useful and meaningful manner. However, the nature of these methodologies is, at best, semi-automated with

1 In computational complexity theory, a problem that is solvable in polynomial time by a non-deterministic Turing machine is an NP (non-deterministic polynomial time) problem. NP-hard problems are those which are at least as hard as any NP-problem, e.g. given a set of integers, does any non-empty subset of them add up to zero?

2 International Virtual Observatory Alliance Knowledge Discovery in Databases Interest Group: http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaKDD.
focused application in particular regions of discovery space rather
than allowing an unbounded exploration of what might be there.

In recent years, a number of approaches have been presented
in the general scientific literature that seek to redress this, e.g.
Oliver et al. (2004), Schmidt & Lipson (2009), Sparkes et al. (2010)
and Reshef et al. (2011). Discovery-based science employs cutting
edge data mining techniques for automated hypothesis forming and
automated theorem proving. Many of these tend to have originated
within the context of systems biology (out of association analysis)
where researchers are attempting to identify and derive universal
relationships in biological systems akin to those which seem to exist
in physical ones, although there is prior art in computer science,
particularly within the area of genetic programming (Koza 1992).

These methods are also similar in scope to various feature se-
lection and extraction and dimensional reduction techniques, such
as principal component analysis (e.g. Francis et al. 1992) and self-
organizing maps (Kohonen 1982), which attempt to counter the
‘curse of dimensionality’ by reducing high-dimensional data to
lower more manageable dimensions whilst preserving meaning-
ful structures within them. Nonetheless, these so-called automated
discovery methods are applicable to any general data set and espe-
cially to those with many variables, such as those which arise in
economics, climate science, sensor networks or any field advocating
an informatics-based approach.

In this work, we describe the application of automated discovery
systems of relationships to astronomical data. We have focused in
particular on two types of approach – those that seek to identify
general connections (correlations) between particular parameters
in a data set and those that try to formulate a specific functional
relationship between parameters. These may be considered repre-
sentative of the type of mapping of discovery space that has so far
been attempted. A common complaint of data mining techniques
is that they usually follow a ‘black box’ approach – the data go in
and the answer comes out but there is no real understanding of how
one led to the other. We hope to show that automated discovery
systems are also more translucent if not actually transparent and
allow some deconstruction of the methodology to understand what
is going on inside. This is particularly important if their discoveries
are to be scientifically validated, i.e. a particular relationship is not
only statistically significant but also stems from a (new) non-trivial
underlying cause.

It should be noted that although these discovery tools are labelled
as automated, they are actually employed as part of a collaborative
human–machine discovery process. In data-intensive problems, not
only are data processing and analysis automated but also necessarily,
given the data volumes and dimensions, the first levels of data
interpretation. The human expert now validates machine-generated
hypotheses rather than attempting to formulate them themselves.
We still make discoveries, but as the complexity of data increases,
we need machine intelligence to help us guide towards an insight.

This paper is structured as follows: in Section 2, we will describe
the two specific techniques we are applying whilst in Section 3, we
will present a number of different astronomical contexts in which
these have been applied – these both attempt to mimic or recreate
past discoveries as well as find new ones. We will analyse and
discuss our results in Sections 4 and 5 and present our conclusions
in Section 6.

2 AUTOMATED DISCOVERY SYSTEMS

The methods we are applying in this paper will probably be unfa-
miliar to many astronomers and so, in this section, we will introduce
some of the pertinent terminology and formalism related to them.

2.1 Maximal information coefficient

The maximal information coefficient (MIC; Reshef et al. 2011) aims
to be the 21st-century equivalent of the Pearson correlation
coefficient (Speed 2011) but goes beyond just expressing linear
associations and can quantify general associations between vari-
ables. It is based on the mutual information between two random
variables, A and B:

\[ MI(A, B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left( \frac{p(a, b)}{p(a)p(b)} \right), \]

where \( p(a) \) and \( p(b) \) are the marginal probability mass functions of \( A \)
and \( B \) and \( p(a, b) \) is the joint probability mass function, respectively.

Now consider a partitioning of a data set, \( D \), of ordered pairs,
\( \{(a_i, b_i), i = 1, \ldots, n\} \), into an \( x \)-by-\( y \) grid, \( G \), such that there
are \( x \) bins (of variable size) covering \( a \) and \( y \) bins (also of variable
size) spanning \( b \), respectively. The probability mass function of
a particular grid cell is clearly proportional to the number of data
points falling inside that cell and so, for a given \((x, y)\), there will be
a maximal mutual information. We can then construct a characteris-
tic matrix \( M(D) \) whose elements:

\[ M(D)_{x,y} = \max_{xy < C(n)} \{ MI(D)_{x,y} \} , \]

are the highest normalized mutual information achieved by any of
the corresponding \( x \)-by-\( y \) grids. The MIC is then defined to be the
maximum value in \( M \), such that \( xy \leq C(n) \), where \( C(n) \) is a function
of the sample size and represents the maximal grid size considered.
Too high a value for \( C \) can lead to non-zero scores even for random
data because each data point gets its own cell, while setting it too
low means that only simple patterns are considered. Reshef et al. (2011)
found empirically that \( C = n^{0.5} \) provides a satisfactory limit:

\[ M(C) = \max_{xy < C(n)} \{ MI(D)_{x,y} \} . \]

The behaviour of MIC is that it tends to 1 for all never-constant
noiseless functional relationships and to 0 for statistically indepen-
dent variables. Moreover, MIC–\( r^2 \), where \( r \) is the Pearson correla-
tion coefficient, is an indicator of a non-linear relationship between
two variables: as \( r \) is a measure of linear dependence, the statistic
MIC–\( r^2 \), is near to 0 for linear relationships and large for non-linear
relationships with high values of MIC. Other measures involving
MIC and \( M \) (the characteristic matrix) can also indicate deviations
from monotonicity, the degree to which the data set appears to be
sampled from a continuous function and the complexity of the asso-
ciation, as different relationship types give rise to characteristic
matrices with different properties.

The statistical significance of an MIC value can be determined
from comparison of a real value against a set of values from \( 1/\alpha \) –
1 surrogate data sets where \( \alpha \) is the probability of false rejection.
Because MIC is a rank-order statistic, the uncorrected \( p \)-value [es-
tentially the one-tailed \( p \)-value for this statistic; when multiple hy-
potheses (many parameters) are being tested, a corrected value
must be used to mitigate false positives] of a given MIC score under a
null hypothesis of statistical independence depends only on the
score and on the sample size of the relationship in question and not
on the specific relationship being tested. Pre-computed uncorrected
\( p \)-values are available for different sample sizes.\(^3\)

To illustrate this statistic, consider a data set of 100 points ran-
domly selected from a cubic relationship \( y = 2x^3 - 3x^2 - 3x + 2\),

\(^3\)http://www.exploredatalab.net/Downloads/P-Value-Tables
2.2 Symbolic regression

Symbolic regression is the task of finding a function, in symbolic form, that fits a finite sample of data. The most efficient approach employs a genetic algorithm-based search (Koza 1992) of the space of mathematical expressions to determine the best-fitting functional form. Successive generations of formulae are specified in terms of a (user-defined) mathematical alphabet of atomic building blocks, such as algebraic and Boolean operators, analytical function types – trigonometric, exponential/logarithmic, power laws, etc., and state variables, which keeps the search tenable. Its advantage over more traditional regression methods is that the search process works simultaneously on both the model specification problem (the form of the fitting equation) and the problem of fitting coefficients.

**EUREQA**

(EUREQA, now also called FORMULIZE) (Schmidt & Lipson 2009) is a software tool which employs symbolic regression to describe a data set by identifying the simplest mathematical formulae which could describe the underlying mechanisms that produced the data. The tool works from the numerical partial derivatives of each pair of variables in the input data set and uses an evolutionary algorithm to explore this partial differential metric space for non-trivial invariant quantities, looking for predicted partial derivatives that best match the numerical ones:

\[
\frac{\Delta y}{\Delta x} \approx \frac{\delta y}{\delta x} \bigg| _{(x_i, y_i)} = \frac{\delta f}{\delta x} \frac{\delta f}{\delta y},
\]

where \( f(x_i, y_i) \) is one of the candidate functions. The search continues until some stopping criterion – time elapsed, goodness of fit, confidence of fit (maturity and stability), etc. – is met. The output is then an ordered list of final candidate analytical expressions on the accuracy–parsimony Pareto front, i.e. the tradeoff between the most optimal (best fit according to some criteria) and complexity. Each mathematical operation in an expression has a numerical value (cost) associated with it, e.g. addition = 1, exponentiation = 4, and the complexity of a formula is defined here as the sum of these values. A high-order polynomial could therefore be more complex than a straightforward exponential or trigonometric function.

When comparing and optimizing solutions, **EUREQA** employs a user-defined error metric. A number of different measures are available and the nature of the data can help determine which is the most appropriate, for example, minimizing the mean of the squared residual errors is suitable for normally distributed noise whereas the logarithmic error is better for many outliers. Data can also be weighted according to some prescription, although the importance of particular variables can always be explicitly stressed in the definition of the equation form being searched for. There are, too, various types of data pre-processing operations available, familiar to data mining, such as normalization, outlier rejection and missing value handling.

The results of symbolic regression, i.e. the expressions identified by **EUREQA**, are the best (non-trivial) mathematical descriptions of the data. Their interpretation and physical validity, however, remain an exercise for the human expert, who may take them at face value or decide to cross-check them (‘prove them’) using other techniques.

3 EXPERIMENTS

In this section, we report on a number of automated discovery experiments we have carried out with different representative astronomical data sets. A number of different options are available, depending exactly on how you want to measure the whole process. It should be noted that in applying our techniques, we are not simply fitting a set of formulae to data but that the respective discovery methods decide which variables to use and in what functional relationship and then find the optimal coefficients and measures of fit. The two methods are also sufficiently different that it is interesting to compare their findings relative to each other.

3.1 The Hertzsprung–Russell diagram

The Hertzsprung–Russell (HR) diagram is the quintessential representation of physical relationships associated with different stages of stellar evolution. The original plot of magnitude versus temperature can be considered as the two-dimensional PDF, \( P(M, T_{\text{eff}}) \); more modern versions also incorporate metallicity and surface gravity giving a four-dimensional PDF, \( P(M, T_{\text{eff}}, [M/H], \log g) \) – which constrains all of its arguments. The parametrization of these relationships in terms of observable and non-observable stellar quantities expressed as a function of the observable colour is an open problem in astronomy, e.g. Wilson & Hurley (2003) and Zaninetti (2008). This is particularly relevant for the next generation of large photometric surveys, e.g. LSST, where spectroscopy of every source
is not feasible. Note that Liu et al. (2012) describe a related problem of inferring the astrophysical parameters of stars from Gaia spectrophotometry.

Unfortunately, prior to the availability of the Gaia data, there is no single large stellar data set which offers both accurate distances and physical parameters for a representative sampling of the HR space. Hipparcos has reliable distances but no intrinsic parameters, such as \( T_{\text{eff}} \) or \([M/H]\). RAVE DR3 (Siebert et al. 2011) and SEGUE (Yanny et al. 2009) both offer spectroscopically determined parameters \((T_{\text{eff}}, g, [M/H])\) but lack distance information – RAVE DR3 shares only 685 objects with Hipparcos and with SEGUE none. A photometric parallax relationship has been defined for SEGUE based on stellar metallicity and colour (Ivezic et al. 2008) but the corresponding HR diagram shows only a main sequence (see Fig. 2).

For a relatively complete coverage of the parameter space, we have therefore constructed a data set consisting of all stars in SIMBAD with a quoted parallax, effective temperature \((T_{\text{eff}})\), surface gravity \((g)\) and metallicity \([M/H])\). The code to look for formulae of the form

\[
M_V = f(B - V, g, T_{\text{eff}}, [M/H])
\]

in the first case (general HR) and

\[
M_V = f(B - V, g, T_{\text{eff}}, [m/H], [\alpha/Fe], \xi, V_{\text{rot}})
\]

in the second (47 Tuc). Although these formulations are based on prior knowledge of what the dependent variables are and also what data are available, symbolic regression incorporates feature selection and so will only use a subset of the most relevant variables, in this case those which persist in successive generations of calculations in the evolutionary algorithm, rather than all available variables (see Section 3.3 for an explicit demonstration of this). We also consider the choice of variables more in Section 4.

We use a set of mathematical building blocks restricted to: constants, basic operators (+, −, *, /), exp(), log(), \(x^2\). We employed an \(R^2\) goodness-of-fit error metric – EUREQA attempts to maximize this quantity in its fits – and selected an 80:20 split of the data in terms of test set and validation set. Data with any missing values were ignored (other options are available) and no weighting was used for any parameter in terms of its error, as the heterogeneity of the data means that not every value has an error associated with it. 27 CPU-hour runs (taking 1.5 h on 18 cores) produced a number of formulae of varying complexity and correlation coefficients of around 0.85 red for both data sets (see Table 1). We also ran it for the SIMBAD data set restricting the formulae to just power-law expressions [no exp() or log() operators].

We shall consider the results obtained for the SIMBAD data set first. To validate the results and test against overfitting, i.e. the
formulae are actually describing random errors or noise in the data instead of any underlying relationship, we determined the median absolute error for each formula when applied to the RAVE DR3 and SEGUE data sets mentioned above. Johnson B and V magnitudes were derived from the SEGUE data using the transformation equations of Lupton (2005) and an absolute V magnitude determined using the inferred parallax relating apparent r magnitude and absolute r magnitude calculated using the photometric parallax method of Ivezić et al. (2008). Any systematic errors that these transformations may introduce can be estimated from a plot of $M_V$ versus $T_{\text{eff}}$ for the SEGUE data compared to the SIMBAD data. Since $T_{\text{eff}}$ is calculated spectroscopically, any photometric offset will show in the relative positions of the main sequences of the two data sets.

The left plot in Fig. 5 shows good agreement between the SIMBAD and RAVE DR3 data sets with an offset of the SEGUE data relative to the other two. This offset can be estimated from the difference between linear fits to the main sequences of both SIMBAD and SEGUE data sets defined between the regions of $T_{\text{eff}} = 5000$ and $T_{\text{eff}} = 6500$ and we find a value of $\Delta M_V = -1.112$ for SEGUE. A similar procedure can be performed with plots of $(B - V)$ versus $T_{\text{eff}}$ to estimate any systematic errors in the colour and we find a value of $\Delta (B - V) = -0.041$ for SEGUE. The right plot in Fig. 5 shows the agreement between the three data sets when the offsets have been applied.

Table 1 gives the median absolute difference (MAD) between the ‘measured’ absolute magnitude and the estimated value for each formula when applied to the SIMBAD, SEGUE and RAVE data sets. For comparison, we also computed the MAD between the observed data and the values derived from the semi-analytical formulae of Zaninetti (2008) relating $M_V$ and $(B - V)$ for each data set (note that Zaninetti’s other formulae relating mass, radius and luminosity to $(B - V)$ all derive from these), although those are only defined over the range $-0.33 < (B - V) < 1.80$ and are stellar luminosity class dependent, with separate relationships for main sequence, giant, supergiant and white dwarf stars.

It is worth bearing in mind when looking at these results that the various functional relationships that this approach finds are, in some statistical sense, the optimal descriptors of the data – they are phenomenological. Their physical interpretation, however, remains the purview of the human scientist. This method aims to identify all potentially interesting, significant relationships without any pre-conceived bias, e.g. due to some established notion of what should actually be there.
The results for the SIMBAD and SEGUE data sets are broadly consistent suggesting that the found formulae provide a good description of the variable relationships in the data but do not overfit it. It should be not surprising that the Zaninetti formulae give better results for the SEGUE data set since it essentially just consists of a main sequence and uses that class specific result. The EUREQA results are derived for a range of luminosity types and so give a better broader fit but not necessarily for specific luminosity classes. The poor performance on the RAVE DR3 data set can be largely attributed to the errors on the parallax value (the mean value is 7.63 mas with a mean error of 1.91 mas) and thus the absolute magnitude (54 per cent of the objects have \( \sigma_{MV} > 1 \)). If we restrict the analysis to those stars with \( \sigma_{MV} < 1 \), we find that the MAD values drop to \( \sim 1 \) for the EUREQA formulae and 0.6 for the Zaninetti formula.

We can also constrain the EUREQA algorithm to use those formulae which contain particular variables or terms: for example, a number of the solutions in both sets of formulae contain a \( g^2 \) term. A set of formulae derived with these limitations has similar MAD values as for the more generic power law.

The relationships found for the 47 Tuc data set are more specific since they only cover post-main-sequence (PMS) stars. Table 2 shows that they fare much better than the Zaninetti formula on this data. We also note that the formula include dependences on parameters related to convection phenomena in stellar atmospheres as would be expected for PMS stars. The metallicities used in the fitting formulae, [M/H], are the uncalibrated ones determined by the RAVE pipeline in Lane et al. (2011) – the uncertainties are 0.1 dex. To compare the fits on PMS stars from the SIMBAD data set, we need to replace [M/H] with an equivalent expression in terms of [M/H]. Zwitter et al. (2008) give a calibration equation for RAVE-derived metallicities:

\[
[M/H] = 0.938[m/H] + 0.767[\alpha/Fe] - 0.064 \log g + 0.404,
\]

but note that \([\alpha/Fe]\) has a typical recovery error of up to 0.15 dex and only spans 0.4 dex in range. Thus, although the SIMBAD data have no measured \([\alpha/Fe]\), we can assume a mean value of 0.2 for use in determining uncalibrated metallicities with reasonable accuracy. We note that the SIMBAD PMS data also show greater intrinsic scatter than the 47 Tuc data.

As Table 3 shows, for the SIMBAD data, MIC identifies statistically significant relationships between \( M_V \) and \((B-V)\), \( T_{eff} \) and \( g \), respectively, but not \([M/H]\). Those involving \( T_{eff} \) and \((B-V)\) are also more likely to be non-linear in nature than that with \( g \). In fact, there seems to be a general set of relationships between \( M_V \), \((B-V)\), \( T_{eff} \) and \( g \) but not with \([M/H]\). Certainly, this is in line with the EUREQA formulae where the \([M/H]\) dependence is not complex but strictly linear. The MIC results for the 47 Tuc data show the significant relationships found in the SIMBAD data set but also ones involving microturbulence and metallicities as we would expect for PMS stars. Note that there is a weak dependence between \( V_{rot} \) and \( M_V \) but not between it and any other parameter. The relationships between \( M_V \) and \((B-V)\), \( T_{eff} \) and \( g \) are also again more likely to be non-linear in nature.

### 3.2 The Fundamental Plane of elliptical galaxies

The global properties of elliptical galaxies, such as luminosity, projected velocity dispersion, etc., form a two-dimensional family (Djorgovski & Davis 1987, hereafter DD87). In particular, an empirical relationship was found by multivariate statistics between the mean surface brightness, central velocity dispersion and effective radius of an elliptical galaxy – the so-called Fundamental Plane – which can be employed as a distance indicator, e.g. Dressler et al. (1987). This has its physical basis in the virial theorem, although there are further structural dependences exhibited between dwarf and giant ellipticals (Guzman, Lucey & Bower 1993).

Using EUREQA, we searched the original data set (161 objects) used by DD87 for relationships of the form

\[
\log(r_v) = f(\log \sigma, \langle \mu \rangle, M),
\]

where \( r_v \) is the semimajor axis, \( \sigma \) is the velocity dispersion, \( \langle \mu \rangle \) is the mean surface brightness and \( M \) is the absolute magnitude in the \( r \)-band. We used a slightly modified set of building blocks from that which we used in the previous section, in which we also allowed for periodic behaviour which could be described in terms of a sine function. This increases the size of search space available, allowing for a wider set of possible relationships, but also, potentially, the computation time. We note, however, that we have no expectation of periodic behaviour; in fact, we know that it would make no sense in this particular context. Rather part of the experiment is just to see what effect allowing for it in the building blocks might have. We also employed a fitness metric based on the mean absolute error.

The best-fitting (lowest complexity, highest accuracy) formula was

\[
\log(r_v) = \log(\sigma) + 0.271(\mu) - 4.09,
\]

which should be compared with the original relationship reported by DD87 (see also Fig. 6):

\[
\log(r_v) = 1.39(\log(\sigma) + 0.26(\mu)) - 6.71.
\]

The two fits have equivalent accuracies – both give rms errors of 0.157 and correlation coefficients of 0.91. Higher order formulae give even slightly better fits, e.g. \( \log(r_v) = (0.14 \log(\sigma) - 1)(\mu) - M \) with 0.142 and 0.92, respectively, but there is a danger that this is overfitting the data, particularly given the small size of the data set. We note also that the sine function was not used.

Given that both relationships are sufficiently similar in form (complexity) and accuracy, the question arises as to which one is correct? This judgement call is beyond the scope of current discovery systems and is where the (human) expert must step in and

---

**Table 2.** The best-fitting formulae found by EUREQA to describe the 47 Tuc HR diagram and the median absolute error obtained when applying the fit to the data.

<table>
<thead>
<tr>
<th>Function ((M_V = f(...)))</th>
<th>Complexity</th>
<th>(R^2)</th>
<th>47 Tuc</th>
<th>SIMBAD PMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.685 \pm \frac{1}{3.19 - MV} - \xi)</td>
<td>8</td>
<td>0.283</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(g) + (\frac{1}{H - 2.39(\alpha/Fe)}) – (\xi)</td>
<td>12</td>
<td>0.260</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(13.2 + (0.0027T_{eff} - 1.62 g - 8.42(B - V))[m/H] - 12.8(B - V))</td>
<td>23</td>
<td>0.239</td>
<td>0.990</td>
<td>0.773</td>
</tr>
<tr>
<td>Zaninetti</td>
<td>–</td>
<td>0.773</td>
<td>0.745</td>
<td>0.773</td>
</tr>
</tbody>
</table>
provide the necessary interpretative knowledge. In this case, the relationships are encoding physical correlations between the size of a galaxy and its effective surface brightness and the luminosity and central velocity dispersion. Even though the form of the expressions is similar, they actually translate into quite different predictions. The DD87 formula gives

$$\langle \mu \rangle \sim L^{-5/4}$$

and

$$D_n \propto \sigma_0^{1.4} \langle \mu \rangle^{-0.07},$$

where $D_n$ is the diameter within which the mean surface brightness is $20.75 \mu_B$ (Dressler et al. 1987), whereas the EUREQA result gives

$$\langle \mu \rangle \sim L^{-3}$$

and $D_n \propto \sigma_0 \langle \mu \rangle^{0.16},$

implying that more luminous galaxies have much lower surface brightnesses and that the distances to galaxies is less than that actually seen.

The EUREQA fit makes no use of $M_V$ and the value of MIC for this variable relative to log ($r_e$) is the lowest, consistent with a lack of dependency. There is also no indication of any type of bivariate relationship beyond a linear one, although EUREQA finds non-linear multivariate relationships to which the MIC is most likely not too sensitive.

### 3.3 Binary classification of light curves

Determining whether an object belongs to a specified class or not, e.g. a transient detection is a supernova (SN) or not, or, alternatively, whether it falls into one of the two different (mutually exclusive) classes, such as star or galaxy, is an increasingly common activity in astronomy (note that multiclass classification problems can always be recast as a series of such binary decisions). This is particularly true of survey astronomy where large data volumes and, most recently, real-time data streams require fast, accurate and reliable classification systems. A variety of techniques have been employed in response, e.g. Djorgovski et al. (2012b), including decision trees (e.g. Vasconcellos et al. 2011), Bayesian networks (e.g. Dubath et al. 2011) and support vector machines (SVM; e.g. Beaumont, Williams & Goodman 2011), the latter representing the state of the art.

Table 3. The MIC measures between the variables used to define the HR diagram for the SIMBAD and 47 Tuc data sets. For these data sets, a value of MIC $>0.41$ for SIMBAD and MIC $>0.17$ for 47 Tuc is significant at the $10^{-4}$ level, respectively. This also illustrates the $n \times (n-1)/2$ nature of the output for a data set of $n$ variables.

<table>
<thead>
<tr>
<th>Variable pair</th>
<th>MIC</th>
<th>Non-linearity</th>
<th>Non-monotonicity</th>
<th>Functionality</th>
<th>Complexity</th>
<th>Linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMBAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>($B - V$) versus $T_{\text{eff}}$</td>
<td>0.82</td>
<td>0.40</td>
<td>0.02</td>
<td>0.82</td>
<td>7.11</td>
<td>$-0.65$</td>
</tr>
<tr>
<td>$M_V$ versus $g$</td>
<td>0.63</td>
<td>0.04</td>
<td>0.05</td>
<td>0.63</td>
<td>7.11</td>
<td>0.77</td>
</tr>
<tr>
<td>$M_V$ versus $T_{\text{eff}}$</td>
<td>0.54</td>
<td>0.48</td>
<td>0.08</td>
<td>0.54</td>
<td>7.11</td>
<td>$-0.24$</td>
</tr>
<tr>
<td>$g$ versus $T_{\text{eff}}$</td>
<td>0.52</td>
<td>0.46</td>
<td>0.03</td>
<td>0.52</td>
<td>7.11</td>
<td>0.24</td>
</tr>
<tr>
<td>$M_V$ versus ($B - V$)</td>
<td>0.49</td>
<td>0.47</td>
<td>0.09</td>
<td>0.48</td>
<td>7.11</td>
<td>$-0.12$</td>
</tr>
<tr>
<td>($B - V$) versus $g$</td>
<td>0.46</td>
<td>0.14</td>
<td>0.0</td>
<td>0.46</td>
<td>7.11</td>
<td>0.14</td>
</tr>
<tr>
<td>$T_{\text{eff}}$ versus $[M/H]$</td>
<td>0.11</td>
<td>0.09</td>
<td>0.02</td>
<td>0.11</td>
<td>7.11</td>
<td>0.09</td>
</tr>
<tr>
<td>$M_V$ versus $[M/H]$</td>
<td>0.09</td>
<td>0.08</td>
<td>0.03</td>
<td>0.09</td>
<td>7.11</td>
<td>$-0.09$</td>
</tr>
<tr>
<td>($B - V$) versus $[M/H]$</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
<td>0.08</td>
<td>7.11</td>
<td>0.01</td>
</tr>
<tr>
<td>$g$ versus $[M/H]$</td>
<td>0.07</td>
<td>0.06</td>
<td>0.02</td>
<td>0.07</td>
<td>7.11</td>
<td>0.10</td>
</tr>
<tr>
<td>47 Tuc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_V$ versus ($B - V$)</td>
<td>0.75</td>
<td>0.54</td>
<td>0.22</td>
<td>0.75</td>
<td>6.43</td>
<td>$-0.45$</td>
</tr>
<tr>
<td>$M_V$ versus $g$</td>
<td>0.62</td>
<td>0.56</td>
<td>0.03</td>
<td>0.62</td>
<td>6.43</td>
<td>0.23</td>
</tr>
<tr>
<td>$g$ versus $T_{\text{eff}}$</td>
<td>0.56</td>
<td>$-0.01$</td>
<td>0.03</td>
<td>0.56</td>
<td>5.75</td>
<td>0.76</td>
</tr>
<tr>
<td>($B - V$) versus $T_{\text{eff}}$</td>
<td>0.56</td>
<td>0.08</td>
<td>0.04</td>
<td>0.56</td>
<td>6.29</td>
<td>$-0.69$</td>
</tr>
<tr>
<td>($B - V$) versus $g$</td>
<td>0.54</td>
<td>$-0.03$</td>
<td>0.06</td>
<td>0.54</td>
<td>6.13</td>
<td>$-0.75$</td>
</tr>
<tr>
<td>$M_V$ versus $T_{\text{eff}}$</td>
<td>0.50</td>
<td>0.49</td>
<td>0.09</td>
<td>0.50</td>
<td>6.43</td>
<td>0.14</td>
</tr>
<tr>
<td>$M_V$ versus $\xi$</td>
<td>0.38</td>
<td>0.36</td>
<td>0.08</td>
<td>0.38</td>
<td>6.25</td>
<td>$-0.12$</td>
</tr>
<tr>
<td>$[\alpha/Fe]$ versus $\xi$</td>
<td>0.34</td>
<td>0.32</td>
<td>0.11</td>
<td>0.34</td>
<td>6.43</td>
<td>$-0.13$</td>
</tr>
<tr>
<td>$T_{\text{eff}}$ versus $\xi$</td>
<td>0.29</td>
<td>0.28</td>
<td>0.13</td>
<td>0.28</td>
<td>6.43</td>
<td>0.07</td>
</tr>
<tr>
<td>$[m/H]$ versus $[\alpha/Fe]$</td>
<td>0.23</td>
<td>$-0.09$</td>
<td>0.01</td>
<td>0.23</td>
<td>6.36</td>
<td>$-0.57$</td>
</tr>
<tr>
<td>$g$ versus $\xi$</td>
<td>0.23</td>
<td>0.12</td>
<td>0.03</td>
<td>0.23</td>
<td>6.43</td>
<td>$-0.33$</td>
</tr>
<tr>
<td>($B - V$) versus $\xi$</td>
<td>0.22</td>
<td>0.14</td>
<td>0.06</td>
<td>0.22</td>
<td>6.43</td>
<td>0.28</td>
</tr>
<tr>
<td>$M_V$ versus $[m/H]$</td>
<td>0.21</td>
<td>0.20</td>
<td>0.09</td>
<td>0.21</td>
<td>6.43</td>
<td>0.11</td>
</tr>
<tr>
<td>$M_V$ versus $V_{\text{rot}}$</td>
<td>0.19</td>
<td>0.19</td>
<td>0.14</td>
<td>0.19</td>
<td>6.43</td>
<td>$-0.02$</td>
</tr>
<tr>
<td>$g$ versus $[m/H]$</td>
<td>0.17</td>
<td>0.03</td>
<td>0.02</td>
<td>0.17</td>
<td>6.43</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Although it seems somewhat counterintuitive, automated discovery systems can also be used as binary classifiers. With EUREQA, the ‘trick’ is to formulate the search relationship as
\[
\text{class} = g(f(x_1, x_2, x_3, \ldots, x_n)),
\]
where \(g\) is either the Heaviside step function or the logistic function, which gives a better search gradient and can be used to produce Receiver Operator Character (ROC)\(^5\) curves for the resulting classification. EUREQA finds a best-fitting function, \(f\), to the data that will get mapped to a 0 or a 1, depending on whether it is positively or negatively valued (or lies on either side of a specified threshold value, say 0.5, in the case of the logistic function). In other words, it finds an equation for the discriminating hyperplane which separates the two classes in some high-dimensional feature space. This is comparable to what an SVM\(^6\) does but with an explicit computation of the mapping into feature space rather than just relying on inner products within it. An advantage of this approach is that the structure of the analytical fit function can also give insight into how the classification works, which is not normally true of other ‘black box’ classifiers, such as neural networks.

The Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009; Mahabal et al. 2011; Djorgovski et al. 2012a) is the largest open-time domain survey currently operating, covering \(\sim 33,000\) deg\(^2\) between \(-75^\circ < \text{ Dec.} < 75^\circ\) (except for within \(\sim 10–15^\circ\) of the Galactic plane) to \(\sim 20\) mag. Light curves of several hundred million objects are available\(^7\) with an average of \(\sim 250\) observations over a 7-year baseline. A common approach to light-curve classification is to characterize the light curves through extracted features, such as moments, flux and shape ratios, variability indices and periodicity measures. Vectors of such features derived from the light curves of known classes of objects are then used as the training sets for particular classifiers.

We have considered three specific binary light-curve classification problems using EUREQA: RR Lyrae versus W UMa, CV versus blazar and Type Ia versus core-collapse SN (CCSN). For each case, we compiled data sets of light curves of the appropriate classes of object and derived \(\sim 30–60\)-dimensional feature vectors for each object (see Appendix 1 for the full list of features used). These are a combination of the features used by Richards et al. (2011) and Debusscher et al. (2007) and include statistical moments, flux ratios, Stetson J and K variability indices, a quasar-fitting measure for both classes of objects are shown in Fig. 7. Since both classes of object are periodic, we included periodic features in our characterization and used 60-dimensional feature vectors.

We ran a set of 10 4-CPU-hour EUREQA runs (1 h on a quad-core machine) for each of the three cases with each run omitting 10 per cent of the data (giving training sets that are 90 per cent of the data set) and the best-fitting solution for that run (defined as the least complex which produces the largest number of true positive and negative class attributions), and then applied with the omitted data as the validation set so giving us 10\(\times\) cross-validation. The results are the sums of each cross-validation run.

Table 4. The combined best-fitting confusion matrices for the three binary classification cases using EUREQA and 10\(\times\) cross-validation. The results are the sums of each cross-validation run.

<table>
<thead>
<tr>
<th></th>
<th>RR Lyrae</th>
<th>W UMa</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR Lyrae</td>
<td>464 (96.3 per cent)</td>
<td>18 (3.7 per cent)</td>
</tr>
<tr>
<td>W UMa</td>
<td>7 (2.5 per cent)</td>
<td>456 (98.5 per cent)</td>
</tr>
<tr>
<td>CV</td>
<td>368 (91.1 per cent)</td>
<td>36 (8.9 per cent)</td>
</tr>
<tr>
<td>Blazar</td>
<td>45 (37.5 per cent)</td>
<td>75 (62.5 per cent)</td>
</tr>
<tr>
<td>SNe Ia</td>
<td>773 (92.5 per cent)</td>
<td>63 (7.5 per cent)</td>
</tr>
<tr>
<td>CC SNe</td>
<td>250 (58.6 per cent)</td>
<td>177 (41.4 per cent)</td>
</tr>
</tbody>
</table>

Eclipsing binaries (W UMa) are the predominant contaminant in studies using RR Lyrae as tracers of Galactic structures, e.g. Sesar (2011), and therefore being able to distinguish between them would be useful. We extracted CRTS light curves for 482 RR Lyrae and 463 W UMa from SIMBAD and the AAVSO International Variable Star Index (Watson, Henden & Price 2006) obtained from VizieR (Ochsenbein, Bauer & Marcout 2000). The magnitude distribution for both classes of objects are shown in Fig. 7. Since both classes of object are periodic, we included periodic features in our characterization and used 60-dimensional feature vectors.

The overall best-fitting formula was
\[
f = 278 x_{24} - \frac{6.63}{x_{10}} - 24,
\]
where \(x_{24}\) is the principal period from the Lomb–Scargle periodogram (Lomb 1976; Scargle 1982) and \(x_{10}\) is the median absolute deviation. The resulting values of class are 1 for RR Lyrae and 0 for W UMa objects. The combined confusion matrix for the best-fitting classification formulae, i.e. summing the individual cross-validation results, is shown in Table 4 and the ROC curve showing the dependences between the true and false positive classification rates, respectively, as the logistic function threshold value is varied in Fig. 9a. It is interesting to note that this is essentially the period–amplitude relation which is used to differentiate between subclasses of RR Lyrae (e.g. Smith, Catelan & Kuehn 2011). Fig. 8 shows how the two populations are clearly separated in this parameter plane.

---

5 An ROC curve is a graphical plot which summarizes the performance of a classifier over a range of tradeoffs between true positive and false positive errors rates (see Fig. 9.)

6 A support vector machine (SVM) is the state-of-the-art binary classification algorithm.

7 http://crts.caltech.edu
Machine-assisted discovery of relationships in astronomy

MIC measures were calculated for all pairs of features in this feature set. We would expect that significant relationships would be found for pairs of variables having a common basis, e.g. those derived from the Lomb–Scargle periodograms of the light curves or those which measure the fraction of outliers or degree of spread in the light curve, and this was confirmed. The MIC measure also largely correlated with the regression coefficient for these pairs, i.e. those with a high MIC value had a high $r^2$ value as well and vice versa, but one strongly related pair (MIC close to 1) had a very low linear regression ($\sim 0.12$). The non-linearity MIC statistic indicated was also large for this pair and the two features were found to be inversely proportional to each other. This clearly illustrates the power of MIC over traditional bivariate relationship analysis algorithms.

MIC analysis of this feature set – calculating the MIC measures for all pairs of features – showed significant relationships between expected pairs of variables, e.g. those derived from the Lomb–Scargle periodograms of the light curves or those which measure the fraction of outliers or degree of spread in the light curve. These largely correlated with the regression coefficient for these pairs but one strongly related pair had a very low linear regression ($\sim 0.12$). The non-linearity MIC statistic indicated such a relationship and the two features were found to be inversely proportional to each other. This clearly illustrates the power of MIC over traditional bivariate relationship analysis algorithms.

Looking for relationships between the class variable for the data set and the features showed a number of significant ($p < 10^{-4}$) pairings. All associations were also deemed to be complex and, with the exception of the median absolute deviation, non-linear. We will discuss these in further detail later in this paper.

3.3.2 CV versus blazar

The light curves of cataclysmic variables (CVs) and blazars can be difficult to differentiate as both exhibit aperiodic/quasi-periodic variability with significant (several magnitudes) sudden outbursts. We extracted CRTS light curves for 404 known CVs\(^8\) and 120 Fermi and MOJAVE blazars.\(^9\) Periodic features were omitted in the characterization, giving 25-dimensional feature vectors. The overall best-fitting formula was

$$f = 140,067 x_{17} \sin \left(\frac{-0.979}{x_1 - 1.481}\right) - 264,152,$$

where $x_1$ is the amplitude and $x_{17}$ is the significance of the $\chi^2$ quasar statistic (Butler & Bloom 2011). The combined confusion matrix for the best-fitting classifying formulae is shown in Table 4b and the ROC curve in Fig. 9b. From the matrix, the classifier is clearly more successful at identifying CVs than blazars. This may reflect stronger class localization for CVs in the feature space than for blazars, i.e. the distribution of CVs in the feature space is more compact and therefore a discriminating (bounding) hyperplane is more easily defined than for blazars. However, it is more likely due to the 10:3 population ratio of CVs and blazars in the data set and a learning bias – the so-called test distribution effect (Weiss & Provost 2003) – that this has created in the classifier, i.e. with more exposure to CVs, the algorithm has preferentially evolved to classify them. We defer further discussion of this issue to Section 5.

MIC analysis of the feature set again shows a number of expected significant relationships, i.e. flux ratios and quasar statistics, although the correlation with the respective regression coefficients is much less than in the RR Lyrae versus W UMa case, which may be related to the lack of periodic features. The relationships also tend to be non-linear but monotonic. In terms of associations with the class variable, only three significant features were found with no clear indication of non-linearity or non-monotonicity.

3.3.3 SN Ia versus CCSN

Spectroscopic confrmation of SNe candidates can be resource intensive and becomes intractable with the increasingly large numbers expected from the next generation of wide-field surveys, e.g. a few hundred thousand from Pan-STARRS and LSST. The Supernova Photometric Classification Challenge (SPCC; Kessler et al. 2010) aimed to improve the state of the art of SN classification algorithms based solely on photometric data, and, in particular, separating out SNe Type Ia, which is important for cosmological studies. We tested our methodology on a set of 836 SNe Ia and 427 CCSNe (Ib, Ic, IIn, IIp) light curves from the SPCC data set. Again, since we do not believe these to be periodic, we used only non-periodic features to characterize the light curves, giving 25-dimensional feature vectors. The overall best-fitting formula was

$$f = \frac{x_{18} - 22.9}{x_{15} + 0.21x_{18}} - x_{13} - x_{10},$$

where $x_{10}$ is the median absolute deviation, $x_{13}$ is the percentage difference between the extremum flux and the median, $x_{15}$ is the $\chi^2$ quasar statistic and $x_{18}$ is the significance of the $\chi^2$ non-quasar statistic, respectively. The combined confusion matrix for the best-fitting classifying formulae is shown in Table 4c and the corresponding ROC curve in Fig. 9c. The matrix again shows a strong classification bias for the more numerous class, although this time the population ratio is only $\sim 2.1$.

The MIC results show expected relationships between flux ratios and measures of variability, all of which are mainly linear, monotonic and in line with the respective regression coefficient results. More interestingly, though, is that there are no really significant

---

\(^8\) http://nesssi.cacr.caltech.edu/catalina/CVservice/CVtable.html
\(^9\) http://nesssi.cacr.caltech.edu/catalina/Blazars/Blazar.html
associations between the class variable and the features – the most significant, the ratio between the (95th–5th) flux percentile and the median, is only significant at the ~3 per cent level. This may indicate that the conventional set of features used to characterize light curves are inappropriate for those of SNe, which would also explain the poor performance of the classifier – a clear discriminating hyperplane cannot be defined in this feature space.

4 FEATURE SELECTION

4.1 Posterior feature selection

In examining data mining systems, it is often worth asking whether a successful outcome is due to the power of the particular algorithm under consideration or due to a comprehensive training data set being used, with which any algorithm would achieve good results. (Alternatively, a poor result with an otherwise excellent algorithm may be due to a limited training set.) One way to answer this is to consider which features in the data set are employed by the algorithm and ask whether the features selected show any degree of sense – do they provide additional insight into the data set – or should we regard them purely as phenomenological selections that just happen to give good results? This is particularly so when only a subset of all the available features actually end up being used, i.e. there is some degree of feature selection present in the process, whether explicit or implicit (embedded), as happens with evolution-based algorithms and the decision tree work we have compared EUREQA against.

The MIC statistics already give some handle on the relative importance of different bivariate relationships within the feature space and of particular features relative to the class variable in the classification examples. However, we would also like to be able to consider larger multivariate subsets of features, both for feature ranking according to some metric and to identify the optimal subset of features that characterizes the problem. We have considered two further specific feature-selection techniques to compare against the results of EUREQA, MIC and the decision trees, and determine whether there is any consistency in the features used by the different techniques: consensus would imply that the shared features are relevant to understanding the problem under consideration.

4.1.1 Sequential backward ranking

Sequential backward ranking (SBR) is an unsupervised feature-selection method based on the entropy measure that aims to progressively reduce the dimension of a data set in an optimal fashion, i.e. at each stage, the reduced data set represents the best approximation to the original. It thus works as follows:

(i) Start with a full feature set \( \mathcal{F} \) which characterizes a data set.
(ii) For each feature, \( f \in \mathcal{F} \), define a set of subsets, \( \{\mathcal{F}_f\} \), such that \( \mathcal{F}_f = \mathcal{F} - f \).
(iii) Select the feature \( f_m \) which maximizes the quantity \( S(\mathcal{F}) - S(\mathcal{F}_{f_m}) \), where \( S(\mathcal{F}) \) is the Shannon entropy (see below) of the feature set \( \mathcal{F} \).
(iv) Update \( \mathcal{F} \) such that \( \mathcal{F} = \mathcal{F} - f_m \).
(v) Repeat steps (ii)–(iv) until there is only one feature left.

The output is an ordered list of features in descending order of their entropy contribution or their significance. A supervised version can also be constructed by replacing the constraint in step (iii) with minimizing the classification error between that for \( \mathcal{F} \) and \( \mathcal{F}_{f_m} \).

In order to apply this technique, we must first define and evaluate the Shannon entropy of a feature set. Traditional estimators of the Shannon entropy, \( H(X) \) of a multivariate data set, \( X = \{X_1, X_2, \ldots, X_n\} \), require knowledge of the joint probability distribution of all the \( X_i \);

\[
H(X_1, \ldots, X_n) = -\sum_{x_1} \ldots \sum_{x_n} P(x_1, \ldots, x_n) \log [P(x_1, \ldots, x_n)]
\]

which is usually a fairly intractable problem. However, Kozachenko & Leonenko (1987) provide an alternative estimator based on the distance to the kth-nearest neighbour:

\[
H(X_1, \ldots, X_n) = -\psi(k) + \psi(n) + \log c_d + \frac{d}{n} \sum_{i=1}^{n} \log(\epsilon_i),
\]

where \( \psi \) is the digamma function \( (\psi(x) = \Gamma'(x)/\Gamma(x)) \), \( c_d \) is the volume of the \( d \)-dimensional unit ball \( (c_d = \pi^{d/2} / \Gamma(1 + d/2)) \) and \( \epsilon_i \) is twice the distance from \( x_i \) to its \( k \)th-nearest neighbour, respectively. The error on the estimate is typically \( \sim k/N \) or \( \sim k/N \log (N/k) \).

4.1.2 Minimum redundancy maximum relevance

In feature selection, it has been recognized that the combinations of individually good features do not necessarily lead to good overall performance, i.e. the \( m \) best features are not the best \( m \) features (e.g. Cover 1974). One way to tackle this is to consider simultaneously the relevance – the average mutual information between a set of features and a classification variable – and the redundancy – the average mutual information between pairs of features – of a feature
Table 5. The features selected by the various feature-selection methods discussed in the text. For MIC, SBR and mRMR, the lists are in descending order of feature significance (indicated by an asterisk). The features to which the variables correspond are described in Table A1.

<table>
<thead>
<tr>
<th>Method</th>
<th>RR Lyrae/W UMa</th>
<th>CV / Blazar</th>
<th>SN Ia/CC SNe</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIC*</td>
<td>x24, x26, x34, x14, x28, x37, x19, x39</td>
<td>x17, x22, x21</td>
<td>–</td>
</tr>
<tr>
<td>EUREQA</td>
<td>x10, x24</td>
<td>x1, x17</td>
<td>x13, x15, x18</td>
</tr>
<tr>
<td>Decision tree</td>
<td>x10, x14, x22, x21, x35, x62</td>
<td>x1, x2, x6, x9, x15, x22</td>
<td>x5, x10, x12, x13, x14</td>
</tr>
<tr>
<td>SBR*</td>
<td>x9, x32, x43, x20, x32</td>
<td>x18, x16, x15, x22, x20</td>
<td>x22, x18, x16, x15, x39</td>
</tr>
<tr>
<td>mRMR*</td>
<td>x9, x20, x11, x19, x13</td>
<td>x2, x17, x12, x17</td>
<td>x1, x2, x22, x15, x38</td>
</tr>
</tbody>
</table>

Table 6. The relative fractions of shared features between those identified by EUREQA or decision trees and those that the three ordered algorithms have provided.

<table>
<thead>
<tr>
<th>Method (MIC/SBR/mRMR)</th>
<th>RR Lyrae/W UMa (8/5/5)</th>
<th>CV / Blazar (3/5/5)</th>
<th>SN Ia/CC SNe (0/5/5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUREQA</td>
<td>50 per cent/0 per cent/0 per cent</td>
<td>50 per cent/0 per cent/50 per cent</td>
<td>0 per cent/67 per cent/33 per cent</td>
</tr>
<tr>
<td>Decision tree</td>
<td>16 per cent/0 per cent/0 per cent</td>
<td>14 per cent/28 per cent/14 per cent</td>
<td>0 per cent/0 per cent/0 per cent</td>
</tr>
</tbody>
</table>

set. Peng, Long & Ding (2005) proposed such a criterion [minimum redundancy maximum relevance (mRMR)]:

$$\max_s \left[ \frac{1}{|S|} \sum_{f_i \in S} \text{MI}(f_i; c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} \text{MI}(f_i; f_j) \right],$$

where the feature set $S$ has individual features $f_i$, $c$ is the classification variable and $\text{MI}$ the mutual information (see equation 1), respectively. This approximates maximizing the mutual information between the joint distribution of the selected features and the classification variable but in terms of bivariate quantities and not much harder to deal with multivariate ones. Note that mRMR employs data discretization as a pre-processing step for continuous data since it is often difficult to compute the integral form of equation (1) in a continuous space with limited numbers of samples.

4.1.3 Feature comparison

Table 5 gives the (ordered) lists of features selected by each method – MIC, EUREQA, decision tree, SBR and mRMR – for the three data sets used in the binary classification problems. We did not consider the HR diagram or Fundamental Plane of elliptical galaxies data sets since these both involved too few variables to show any differences between the methods. The MIC results are those deemed to be statistically significant at the $10^{-3}$ level relative to the classification variable (see Section 2.1 for details). Similarly, the mRMR results are also relative to the classification variable. Feature data for the mRMR algorithm was discretized into three states at the positions $\mu \pm \sigma$ (where $\mu$ is the mean and $\sigma$ the standard deviation, respectively) such that it takes $-1$ if it is less than $\mu - \sigma$, $1$ if larger than $\mu + \sigma$ and $0$ otherwise. The SBR and mRMR results also just list the top five features in each case. Finally, the EUREQA and decision tree entries list the variables used without any implied ranking.

The disparate nature of the rankings – ordered and unordered, different numbers of variables – makes any formal quantitative analysis, such as ranking aggregation, difficult. However, there a number of general comparisons that can be made. The features employed by EUREQA and decision trees are generally different – they only share one feature in each of the three problems – and decision trees are less parsimonious with more features. A similar lack of commonality is shown between MIC and both SBR and mRMR, although SBR and mRMR show a marginally stronger degree of overlap, which should not be that surprising since they both rely directly on entropy-related measures. Table 6 gives the relative fractions of the features selected by EUREQA and decision trees that are also identified by MIC, SBR and mRMR, respectively. This suggests that there is more agreement between EUREQA and the three explicit feature-selection methods than between decision trees and the same techniques, although none of them display any particularly strong association.

The differing nature of the classes of object in the three experiments leads one to expect that specific features or types of features would be selected in each and this does seem to be the case. The RR Lyrae/W UMa results include periodic measures for all methods except mRMR, reflecting the period–amplitude relationship, and the CV/Blazar results include either the QSO or non-QSO statistic for all methods. More interestingly, the QSO statistics are also selected by EUREQA, SBR and mRMR as discriminating features in the SNe data set (the other features selected are not common across the methods). These statistics measure the applicability of a damped random walk model to a light curve versus it exhibiting temporally uncorrelated variability. Although neither behaviour is shown in either type of SN light curve, both exhibiting a brightening and then decaying pattern with additional features in CCSNe, there must be some further information inherent in the light curves to which these statistics are sensitive.

Graczyk & Eyer (2010) propose that eclipsing binaries can be identified in large photometric surveys based on the skew and kurtosis of their light curves. MIC finds some degree of relationship between the skew and kurtosis but not a non-linear one and a stronger non-linear dependence between the class type and the skew than the kurtosis (the dependence of which is actually not statistically significant). mRMR identifies both skew and kurtosis, however, as significant features. Neither are flagged by EUREQA or decision trees, although in the former case, the ‘survivability’ of the period – MAD solution dominates that of other possible relationships. Restricting EUREQA to non-periodic features gives a set of formulae all dependent on the skew and variously the kurtosis and percentile ratios. A viable EUREQA-based feature-selection strategy, particularly for feature-rich data sets, might therefore be to progressively restrict the set of features that are considered in any single iteration.

It is also worth noting which features are not selected at all or only once by one method: flux ratios and the Stetson K variability index. These statistics can be broadly thought of as quantitative measures...
of the shape of the light curve and there is indeed little discrimina-
tion to be found in the shapes of the three binary categories of object
alone, e.g. although RR Lyrae AB and W UMa are relatively easily
distinguished from their phased light curves, RR Lyrae C are not.
CV and blazar light curves are similarly not easily separable, par-
ticularly when sparsely and irregularly sampled such as the CRTS
light curves. SN Ia and CCSNe can be differentiated if enough of
the light curve has been sampled but this is not necessarily the case
with many of the examples in the SPCC.

This suggests that the current features which aim to capture the
shape of a light curve are neither robust enough in the presence of
noisy inhomogeneous data nor do they capture enough information
to act as significant discriminators. Clearly further research in this
area would be extremely beneficial to the next generation of time
domain surveys.

5 DISCUSSION

The results in the previous sections show that automated discovery
systems of relationships can identify and characterize physically
meaningful structure in data. The fact that known relationships in
the HR diagram, the Fundamental Plane of elliptical galaxies and
the period–amplitude plane of RR Lyrae stars can be automatically
recovered is very encouraging, particularly as more complex and
accurate (with smaller errors) expressions are possible. The fitness
metrics, however, provide a good balance between accuracy and
parsimony, ensuring high-quality general hypotheses. It should also
be noted that the discovery process consists not only of identifying
the best functional expressions but also the most relevant subset
of variables. There are, of course, other specific feature-selection
algorithms, such as those mentioned in Section 4, that could have
been applied to the data sets prior to the application of our methods
as a pre-processing step.

Perhaps one of the more surprising applications of these systems
is as part of efficient binary classifiers, particularly as it has been
said that we should not expect a lightly parametrized form for
mapping between feature space and class space (Richards, private
communication). To get some idea of how competitive this approach
can be, we can compare directly with the results of Donalek et al. (in
preparation) who have applied C4.5 decision trees using the Gini
diversity index as the splitting criterion to the same data sets as used
in Section 3. Table 7 gives the relative performances of the two
approaches in terms of purity – the fraction of true classifications
recovered out of all objects assigned to that class – and efficiency
– the fraction of true classifications recovered out of all objects
actually belonging to that class. For example from Table 4, 464
true RR Lyrae are recovered, 471 (464 + 7) objects are assigned
a class of RR Lyrae, and there are 482 (464 + 18) RR Lyrae in
the data set – this gives a purity of 464/471 (98 per cent) and an
efficiency of 464/482 (96 per cent). It can be seen that for four of
the classes, the EUREQA-based approach performs as well as the
decision tree one, particularly in terms of efficiency (completeness).

For additional comparison, the best results reported in SPCC were
96 per cent efficiency and 79 per cent purity for classifying SNe Ia
(Kessler et al. 2010). However, as expected, it does not perform so
well with the two minority class populations.

Imbalanced data sets, such as the CV/blazar and SN Ia/CC SN
examples, may reflect natural class distributions – one type of object
is just more common than the other – or may be the result of pa-
rameter/feature space sampling – observations are probing regions
preferentially occupied by one class, even if the overall population
sizes are similar. We chose to use as much data as possible in both
the CV/blazar and SN Ia/CC SN cases, which probably involves a
mixture of both these effects. With such data sets, minority class
eXamples are classified incorrectly much more often than majority
class examples (Weiss & Provost 2003), as we found in Section 3.

Determining what the correct distribution is for a learning algorithm
in this context is an active area of research in machine learning (see
Chawla 2010 for an overview). Some practitioners believe that the
naturally occurring marginal class distribution should be used so
that new examples will be classified using a model built from the
same underlying distribution. Others feel that the training set should
contain an increased percentage of minority class examples or the
induced classifier will not classify minority class examples well.

Weiss & Provost (2003) show that the choice of training distribu-
tion can depend on the performance measure used with the natural
distribution for predictive accuracy (confusion matrices) and a bal-
canced distribution for ROC curves, respectively. Although, Chawla
(2010) argues that predictive accuracy may be an inappropriate per-
formance measure for imbalanced data sets. Our results are certainly
consistent with all these findings.

Finally, it is worth considering the limitations of automated dis-
coverey systems. EUREQA assumes that relationships must be express-
able as invariant (conserved) quantities in a partial differential metric
space. However, this would not necessarily be true for systems that
might be exhibiting fractal behaviour, such as scale dependence in
the correlation properties of the large-scale distribution of galaxies
(e.g. Joyce et al. 2005), or chaotic or stochastic activity, such as
in accretion discs (e.g. Karak, Dutta & Mukhopadhyay 2010). It is
computationally expensive and more than linearly so as the size of
the search space is increased with the number of building blocks
employed in search formulae. It can also suffer from the general
limitations of evolutionary algorithms, requiring time to move out
of local minima and the nature of the fitness landscape being un-
clear so it is difficult to determine how well the algorithm might be
doing. Although there is no guarantee that it will find a good solu-
tion nor that this will be the optimum, the results we have shown
demonstrate that it is a useful technique to consider.

MIC and its associated statistics have fewer assumptions in
just looking for and broadly characterizing bivariate relationships
through their maximal mutual information – note that they are not
directly related to mutual information as they perform well in sit-
suations where other direct mutual information-based measures do
not (Reshef et al. 2011). Ideally, the MIC algorithm would optimize
over all possible grid partitions of a data set but the computational
expense is avoided with a dynamic programming approach that ap-
pears to approximate well in most cases. It is unclear, however, how
well these perform in the presence of outliers or how large data
sets need to be for stable estimates. Probably the biggest current
limitation to MIC is its bivariate nature – generalizations to higher

Table 7. The overall success rates for the EUREQA-based classi-
fiers and the decision trees of Donalek et al. (in preparation)

<table>
<thead>
<tr>
<th>Data set</th>
<th>EUREQA Purity (per cent)</th>
<th>EUREQA Efficiency (per cent)</th>
<th>Decision tree Purity (per cent)</th>
<th>Decision tree Efficiency (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR Lyrae</td>
<td>98</td>
<td>96</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>W UMa</td>
<td>97</td>
<td>99</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>CV</td>
<td>89</td>
<td>91</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>Blazar</td>
<td>68</td>
<td>63</td>
<td>87</td>
<td>83</td>
</tr>
<tr>
<td>SN Ia</td>
<td>76</td>
<td>93</td>
<td>90</td>
<td>96</td>
</tr>
<tr>
<td>CC SN</td>
<td>74</td>
<td>41</td>
<td>92</td>
<td>80</td>
</tr>
</tbody>
</table>
dimensions are necessary to search for multivariate relationships (these will not necessarily show in a two-dimensional projection) but this comes at the additional expense of both finding an optimal hypergrid partitioning of the data set and also using multivariate mutual information which is a poorly understood concept.

6 CONCLUSIONS

In this paper, we have demonstrated that automated discovery systems can uncover significant (non-trivial) relationships in high-dimensional complex data parameter spaces. As with any relationship found in data, whether by an automated system or a human, these may or may not have a physical meaning or cause -- correlation does not imply causation -- and they may be due to some incidental properties of a given data set. The interpretation and evaluation of their possible physical significance remains in the hands of a human scientist.

Whilst the ones we have shown may not be the most scientifically exciting, being more for illustrative purposes than anything, we should bear in mind that astronomy is an already relationship-rich science. Many of these are expected or predictable associations, given what we already understand about the nature of (astro)physical systems. In contrast, systems biology and similar sciences, wherein lie the origins of these automated discovery techniques, are relationship-poor and there is potentially more upfront impact to be had by applying them in that particular context. Astronomy perhaps stands to benefit more from them as discovery filters, tackling the curse of dimensionality of high-dimensional parameter spaces and reducing the number of relationships to be examined to only the most significant, than as ab initio discovery engines.

Although these systems represent the cutting-edge of currently applicable tools, this is very much an initial entry point for their application to astronomy. Such tools will very likely become both more powerful and also more prevalent with time, given the data challenges all sciences are facing. Expanded abilities such as not just relying on brute-force searches of feature spaces but being able to incorporate domain knowledge, both as additional features, e.g. distance to nearest galaxy for SNe, or as rulesets, and make inferences leading to more interesting discoveries are active areas of research.

The importance of such approaches as we are faced with ever more parameter/feature rich data sets cannot be underestimated. In particular, the possibilities of high-dimensional scientific relationships, particularly those that do not necessarily reveal themselves in lower dimension representations, can only really be investigated using automated discovery techniques which are (relatively) unconstrained in their exploration of parameter space. These tools promise to cherry pick the higher hanging fruit of LSST, SKA and future surveys.

ACKNOWLEDGEMENTS

We thank Hod Lipson and Michael Schmidt for useful discussions and their kind assistance with the EUREQA software. We also thank the anonymous referee for their useful comments which helped improve this paper.

This work was supported in part by the NSF grants AST-0909182 and IIS-1118041, by the W. M. Keck Institute for Space Studies, and by the US Virtual Astronomical Observatory, itself supported by the NSF grant AST-0834235.

This research has made use of data obtained from or software provided by the US Virtual Astronomical Observatory, which is sponsored by the National Science Foundation and the National Aeronautics and Space Administration.

This research has made use of the SIMBAD data base, operated at CDS, Strasbourg, France, and the International Variable Star Index (VSX) data base, operated at AAVSO, Cambridge, Massachusetts, USA.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation and the U.S. Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

REFERENCES

Djorgovski S. G. et al., 2012a, in Mihara T., Serino M., eds, Special Publ.- IPCR-127: The First Year of MAXI: Monitoring Variable X-ray Sources. RIKEN, Tokyo, p. 263
Kessler R. et al., 2010, PASP, 122, 1415
Kohonen T., 1982, Biol. Cybern., 43, 49
Table A1. This table describes the features used to characterize the light curves used in this analysis.

<table>
<thead>
<tr>
<th>Name</th>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude</td>
<td>$x_1$</td>
<td>Half the difference between the minimum and maximum magnitudes</td>
</tr>
<tr>
<td>Beyond 1 std</td>
<td>$x_2$</td>
<td>The percentage of points beyond one standard deviation from the weighted mean</td>
</tr>
<tr>
<td>Flux percentile ratio (60–40)</td>
<td>$x_3$</td>
<td>The ratio of flux percentiles: $(60^\text{th}–40^\text{th})$ to $(95^\text{th}–5^\text{th})$</td>
</tr>
<tr>
<td>Flux percentile ratio (67.5–32.5)</td>
<td>$x_4$</td>
<td>The ratio of flux percentiles: $(67.5^\text{th}–32.5^\text{th})$ to $(95^\text{th}–5^\text{th})$</td>
</tr>
<tr>
<td>Flux percentile ratio (75–25)</td>
<td>$x_5$</td>
<td>The ratio of flux percentiles: $(75^\text{th}–25^\text{th})$ to $(95^\text{th}–5^\text{th})$</td>
</tr>
<tr>
<td>Flux percentile ratio (82.5–17.5)</td>
<td>$x_6$</td>
<td>The ratio of flux percentiles: $(82.5^\text{th}–17.5^\text{th})$ to $(95^\text{th}–5^\text{th})$</td>
</tr>
<tr>
<td>Flux percentile ratio (90–10)</td>
<td>$x_7$</td>
<td>The ratio of flux percentiles: $(90^\text{th}–10^\text{th})$ to $(95^\text{th}–5^\text{th})$</td>
</tr>
<tr>
<td>Linear trend</td>
<td>$x_8$</td>
<td>The slope of a linear fit to the light curve</td>
</tr>
<tr>
<td>Maximum slope</td>
<td>$x_9$</td>
<td>The maximum absolute flux slope between two consecutive observations</td>
</tr>
<tr>
<td>Median absolute deviation</td>
<td>$x_{10}$</td>
<td>The median discrepancy of the fluxes from the median</td>
</tr>
<tr>
<td>Median buffer range percentage</td>
<td>$x_{11}$</td>
<td>The percentage of fluxes within 10 per cent of the amplitude from the median</td>
</tr>
<tr>
<td>Pair slope trend</td>
<td>$x_{12}$</td>
<td>The percentage of the last 30 pairs of consecutive flux measurements that have a positive slope</td>
</tr>
<tr>
<td>Percent amplitude</td>
<td>$x_{13}$</td>
<td>The largest percentage difference between either the maximum or minimum flux and the median</td>
</tr>
<tr>
<td>Percent difference flux percentile</td>
<td>$x_{14}$</td>
<td>The ratio of the $(95^\text{th}–5^\text{th})$ flux percentile to the median flux</td>
</tr>
<tr>
<td>QSO</td>
<td>$x_{15}–x_{18}$</td>
<td>The chisq/qso and chisq/non-qso statistics and their significance levels from the quasar (non-)variability metric of Butler &amp; Bloom (2011)</td>
</tr>
<tr>
<td>Skew</td>
<td>$x_{19}$</td>
<td>The skew of the magnitudes</td>
</tr>
<tr>
<td>Small kurtosis</td>
<td>$x_{20}$</td>
<td>The kurtosis of the magnitudes</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$x_{21}$</td>
<td>The standard deviation of the magnitudes</td>
</tr>
<tr>
<td>Stetson J</td>
<td>$x_{22}$</td>
<td>The Welch–Stetson J variability index with an exponential weighting scheme</td>
</tr>
<tr>
<td>Stetson K</td>
<td>$x_{23}$</td>
<td>The Welch–Stetson K variability index</td>
</tr>
<tr>
<td>Lomb–Scargle peaks</td>
<td>$x_{24}–x_{33}$</td>
<td>The periods and false-peak detection probabilities of the top five peaks in the Lomb–Scargle periodogram of the light curve</td>
</tr>
<tr>
<td>Frequency parameters</td>
<td>$x_{34}–x_{62}$</td>
<td>The frequency analysis statistics described in Debosscher et al. (2007): the slope of the linear trend, the three prime frequencies and their first four harmonics (amplitude and phase for each) and the ratio of the variances of the light curve after and before subtraction of a harmonic fit with the first frequency.</td>
</tr>
</tbody>
</table>