Estimating Photometric Redshifts with Genetic Algorithms

Nick Miles
Computer Science, University of Kent
Canterbury, Kent, CT2 7NF, UK
nick@terado.co.uk

Alex Freitas
Computer Science, University of Kent
Canterbury, Kent, CT2 7NF, UK
A.A.Freitas@kent.ac.uk

Stephen Serjeant
Physical Sciences, University of Kent
Canterbury, Kent, CT2 7NH, UK
s.serjeant@kent.ac.uk

ABSTRACT
Photometry is used as a cheap and easy way to estimate redshifts of galaxies, which would otherwise require considerable amounts of expensive telescope time. However, the analysis of photometric redshift datasets is a task where it is sometimes difficult to achieve a high classification accuracy. This work presents a custom Genetic Algorithm (GA) for mining the Hubble Deep Field North (HDF-N) datasets to achieve accurate IF-THEN classification rules. This kind of knowledge representation has the advantage of being intuitively comprehensible to the user, facilitating astronomers' interpretation of discovered knowledge. The GA is tested against the state of the art decision tree algorithm C5.0 [6] achieving significantly better results.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *induction, concept learning.*

General Terms: Algorithms, Experimentation, Performance

Keywords: Genetic Algorithms, Data Mining, Classification, Photometric Redshift, Spectroscopic Redshift

1. INTRODUCTION

1.1 Spectroscopy & Photometry
For deep field surveys, spectroscopy can prove too time costly to be worthwhile in identifying high redshift objects [5]. Photometry provides a practical and cost-effective method to obtain comparable results. Photometry is the measurement of fluxes through broad filters of astronomical objects, and from these measurements redshifts can be estimated. Improving on the accuracy of this can allow astronomers to take more reliable measurements without the cost and time spent on the telescopes.

1.2 Contribution
This problem is an interesting candidate for Evolutionary Algorithms (EA) because the data is inherently noisy and EAs in general are robust to noise. This work intends to explore how EAs can be used effectively to predict the correct classes, identified by spectroscopy, with greater accuracy than previously used methods. In addition, this produces comprehensible rules that astronomers can use to gain more insight about the data and the application domain. The discovered rules are expressed in a simple IF-THEN structure, and so they provide knowledge that is intuitively interpretable by astronomers, unlike, for instance, the output of black box classification algorithms such as standard artificial neural networks.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA
ACM 1-59593-186-4/06/0007

2. EXTRAGALACTIC ASTRONOMY
Redshift is the decrease in frequency of the light from when it was emitted, a common effect of universal expansion which causes the distance from other galaxies to increase. The expansion of the universe results in a stretching of space, including of the light rays from when they were emitted at their source to when they arrive to us. This stretching means the light has a longer, redder wavelength. Because of this effect, there is a relationship between distance and the amount of redshift; therefore, redshift is often used as a measure of distance to galaxies.

Redshifts of extragalactic objects can be measured via spectroscopy. Broadband photometry looks at far wider wavelength intervals, requiring a much shorter exposure time. So, comparisons can be made with predictions from galaxy Spectral Energy Distributions to determine the photometric redshift.

3. PREPARING THE DATA TO BE MINED
For the proposed algorithm to perform well requires a comprehensive starting dataset. There are several datasets covering the HDF-N region, however they catalogue a range of objects and resolutions. To build a single catalogue across all magnitudes would require overlapping datasets.

Data was collected from TKRS [8], GOODS [1] and Fernandez-Soto [4]. Merging these catalogues was done based upon the position on the sky, to prevent misclassifications of nearby objects the magnitudes are also compared. A combined single dataset was produced consisting of 4398 records covering a range of wavelengths, including the B, V, R, I and Z bands.

From these magnitudes a set of constructed colour attributes was formed from the differences between the magnitudes (i.e. B-V, V-R, R-I and I-Z). Magnitudes are proportional to the logarithm of photon flux, so the differences between magnitudes are flux ratios. Colour is a measure of the magnitude difference of a star or galaxy in two pass bands, relative to the magnitude difference of Vega in the same pass bands. A colour less than zero indicates a temperature hotter (bluer) than that of Vega (around 10,000K), and higher than zero will be cooler (redder) [7].

4. A GA FOR ESTIMATING REDSHIFTS
Each individual represents a single classification rule of the form IF (conditions) THEN (class). Each rule condition is internally encoded into the individual as a gene. Each gene consists of a sextuplet of elements, defined as <LowerValue, Operator, Attribute, Operator, UpperValue, Active Flag>. The Attribute

element is one of the colour attributes. The two Operator elements are both the relational operator less than or equal to "<=". LowerValue and UpperValue are thresholds representing the lower and upper values of the attribute. Active Flag switches on or off this condition of the rule. The individual encoding contains one gene for each attribute of the data being mined.

Each run of the GA discovers a single rule, so many runs of the GA will be required to discover a set of rules covering all training examples. In order to discover a set of classification rules, we use the sequential covering approach. When using these rules on the test set, for each test example, all rules covering that example are identified. If all rules covering the example predict the same class, the example is simply assigned that class. If multiple rules covering the example predict different classes, the rule with the highest fitness is chosen. Finally, if no rules cover the example, the default rule is chosen, predicting the class which has the most examples in the training set.

The fitness function in the GA measures the predictive accuracy of each individual (candidate classification rule). It first builds a confusion matrix based on the predictive accuracy of the rules against the examples in the training set. Then the confidence (TP/(TP+FP)) and completeness (TP/(TP+FN)) are multiplied to produce the fitness, i.e.: Fitness = Confidence * Completeness, where TP, FP, FN, TN stand for the number of true positives, false positives, false negatives and true negatives.

The crossover method used in our GA is uniform crossover with a probability of 70%. Each individual's gene elements probabilistically undergo mutation. The elements that are affected include the upper threshold, the lower threshold and the active flag with a mutation probability on each of 0.01%. Mutation is applied to the upper and lower thresholds by increasing or decreasing by up to 0.5, and to the active flag by flipping it to set whether this particular gene will be active. We use tournament selection with replacement and a tournament size of 5.

5. RESULTS AND DISCUSSION

In order to evaluate the proposed GA, we compared its performance against the performance of C5.0 [6]. To make a direct comparison with the classification rules discovered by the GA, we used the rule set mode of C5.0. Both the GA and C5.0 were evaluated by running a 10-fold cross-validation procedure [9]. We used the default parameters of C5.0 and of the GA in order to make the comparison with C5.0 as fair as possible. In all runs of the GA, the population size was 100 individuals, and the number of generations was 10. Although this is a small number of generations, by comparison with values typically used in the GA literature, even with this value we found a set of rules representing a significant improvement over the set of rules discovered by C5.0, as discussed below. The results reported here are the average classification accuracy rate over the test set (unseen during training) across the 10 iterations of the cross-validation procedure. In that procedure, the same partition of the data into 10 folds was used by both the GA and C5.0, again to make the comparison between the two systems as fair as possible.

For each of the iterations of the cross validation procedure, the GA was run 10 times with different random seeds, whilst C5.0 was run just once, since it is a deterministic algorithm. The final average classification accuracy across the 100 runs of the GA is (93.16 +/- 0.46)%. The average accuracy of C5.0 over the 10 runs is (90.73 +/- 0.63)%. These results have >99% confidence that they are statistically significantly different and not by chance, as measured by a Student t-test. This can be considered a very good result, considering that C5.0 is the product of several decades of research in decision tree and rule induction, whilst the GA proposed here is still in its first version.

The simplicity (number of discovered rules and terms, or conditions, for all discovered rules) is 8.43 (+/- 1.64) rules and 10.01 (+/- 2.66) terms for the GA, and 16.5 (+/- 2.58) rules and 41.8 (+/- 5.69) terms for C5.0. The measure of simplicity suggests that the rules generated by the GA are considerably simpler, and therefore more easily interpretable, than the rules generated by C5.0. These measurements of simplicity have >99% confidence of being statistically significantly different and not by chance, as measured by a Student t-test.

6. FUTURE RESEARCH

The only attributes used in these runs were the constructed colours attributes. One research direction will be to introduce further attributes, including morphological parameters, by building up a more comprehensive catalogue. Another research direction will be to use relational conditions in the attribute space (e.g. B-V < R-I). Relational conditions are used in astronomy, for example in finding star forming galaxies [3]. A third research direction is to perform template matching using HYPER-Z [2] and to compare the results with the GA.

One of the main problems with using a χ^2 solution, such as HYPER-Z, is that it can sometimes confuse spectral features, such as the Balmer and Lyman breaks, therefore misclassifying. With further attributes we intend to build more sophisticated rules with the GA which will aim to prevent such misclassifications, or identify them as misclassified and correct them.

7. REFERENCES

[1] Cowie L.L. et al. A large sample of spectroscopic redshifts in ACS-GOODS region of the HDF-N, astro-ph/0401354, 2004

[2] Bolnzella, M. et al. *Photometric redshifts based on standard SED fitting procedures*, Astrophys. J., 363, 476–492, 2000

[3] Daddi et al., Star-forming and Passive Galaxies, ApJ, 617, 746, 2004

[4] Fernandez-Soto A. et al. A new catalog of photometric redshifts in the Hubble Deep Field, Astrophys. J., 513, 34-50, 1999

[5] Gwyn, S., The Redshift Distribution and Luminosity Functions, astro-ph/9603149, 1996

[6] Rulequest Research, Data Mining Tools See5 and C5.0 http://www.rulequest.com/see5-info.html, Visited Oct 2005

[7] Richmond, M., Photometric Systems and Colors http://spiff.rit.edu/classes/phys445/lectures/colors/colors.html, Visited Oct 2005

[8] Wirth G. et al. The Team Keck Treasury Redshift Survey of the GOODS-North Field, astro-ph/0401353, 2004

[9] Witten I. H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005