Estimating Photometric Redshifts Using Genetic Algorithms

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

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

Stephen Serjeant
Physical Sciences
Open University,
Milton Keynes, Bucks,
MK7 6AA, 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 [Rulequest, 2005] in two datasets, achieving better classification accuracy and simpler rule sets in both datasets.

1. INTRODUCTION

1.1 Spectroscopy & Photometry

Astronomers today face the challenge of needing large amounts of spectroscopic telescope time in extremely deep field surveys, to identify high redshift objects (where *redshift* is a shift in frequency of the light towards red; see section 2.1). Spectroscopy can prove too time costly to be worthwhile [Gwyn, 1996]. 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.

Current methods for estimating redshift using photometry are unfortunately not perfect. By improving on the accuracy of this one can allow astronomers to take more reliable measurements without the cost and time spent on the telescopes.

1.2 Contribution

This paper proposes an Evolutionary Algorithm (EA) for discovering classification rules – supervised learning problem in machine language terminology. 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, as will be described later in section 4, 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.

2. Extragalactic Astronomy

2.1 What is redshift?

Redshift is the decrease in frequency (towards the red end of the spectrum) of the light from when it was emitted. This is commonly a result of the emitting and receiving locations moving apart from one another, causing the light rays to stretch (conceptually similar to the Doppler shift, also a commonly observed effect with sound waves). Redshift (z) is formally defined as:

$$1 + z = \frac{\lambda_{observed}}{\lambda_{emitted}}$$

where $\lambda_{observed}$ is the observed wavelength of the light, and $\lambda_{emitted}$ is the wavelength emitted at the source.

When observing distant galaxies, the observed shift is not caused by the movement of galaxies away from us. The distance is simply too great to be able to discern such changes. Instead, it is the effect of universal expansion being observed, which in turn is causing 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. The further away the observed astronomical objects are, the larger this effect is, as the light travels a greater distance and therefore undergoes more prolonged stretching.

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.

2.2 Finding high-redshift objects

Distant objects are very faint and small. Finding these against the many closer small and faint objects is often difficult, involving taking spectra to directly observe

the redshift of these. Therefore it can be hard to reliably find and catalogue a large number of high-redshift objects. Inevitably with enough spectra, one will find some of the high-redshift objects. Deep field imaging will therefore find many more faint, distant objects than it is practicable to obtain spectra for.

2.3 Photometric Redshift

Redshifts of extragalactic objects can be measured via spectroscopy. Emission and absorption lines are identified and wavelengths measured, then compared to known rest wavelengths to determine redshift.

Broadband photometry looks at far wider wavelength intervals, typically 1000Å wide, therefore requiring a much shorter exposure time. Comparisons can then be made with predictions from galaxy Spectral Energy Distributions (SED) to determine the photometric redshift (Figure 1). Rather than observing narrow spectral features of galaxy spectra, the photometric redshift technique concentrates on broad features, such as the 4000Å break and the overall shape of a spectrum [Gwyn, 1996].

One method used takes a training set of measured spectroscopic redshifts and derives an empirical relation between these measured redshift and observed magnitudes [Connolly et al., 1995]. Another version of this method derives redshifts by way of a linear function of colours [Wang et al., 1998].

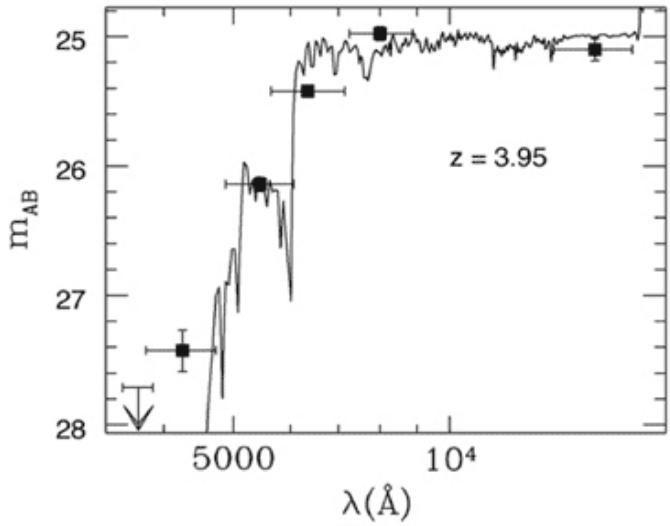

Figure 1: An example of photometric redshift determination for a faint galaxy in the HDF-S NIC3 field. The filled points are the fluxes measured in the five colours observed with the VLT Test Camera (U, B, V, R and I) and in the infrared H spectral band with the NICMOS instrument on the HST. The curves constitute the best fit to the points obtained from a library of more than 400,000 synthetic spectra of galaxies at various redshifts. [ESO, 1998]

2.4 Deep Field Surveys

Deep field surveys are becoming more common: examples include the HDF (Hubble Deep Field) North and HDF South, the NTT Deep Field, and the FORS Deep Field. These were all chosen for the sparseness of highly luminous foreground objects [ESO, 1999].

One of the major motivations currently in deep field surveys is the measurement of colours for all observed galaxies in the field so that photometric redshifts can be inferred.

3. Preparation of the data to be mined

Data was collected from several sources covering the HDF-N including Team Keck Treasury Redshift Survey (TKRS) [Wirth et al., 2004], Hubble's Advanced Camera for Surveys (ACS) Great Observatories Origins Deep Survey (GOODS) [Cowie et al., 2004] and Fernandez-Soto spectroscopic data set [Fernandez-Soto et al., 1999]. Merging these catalogues was done based upon the position on the sky. To prevent misclassifications of nearby objects the magnitudes are also compared. IDL (Interactive Data Language) was used to cross reference and match the entries, which was then used to combine the data into one dataset with just the necessary parameters for the application.

This produced a dataset of 4398 records covering a range of wavelengths, including the B, V, R, I and Z bands. B, V, R, I and Z bands (also referred to as the UBVRI system) are a commonly used set of wideband filters, with a large range in wavelengths of near visible light.

Finally, from these magnitudes two sets of constructed attributes were formed. The first set 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 (or bluer) than that of Vega (around 10,000K), and higher than zero will be cooler (or redder) [Richmond, 2005].

The second set of attributes was formed from the differences between the colours (i.e. (B-V)-(V-R), (V-R)-(R-I) and (R-I)-(I-Z)), this was then added to the colours to improve upon the predictive power of that dataset. The differences between colours is also described as BzK. BzK photometry is a method used for selecting actively star forming galaxies, independently of dust reddening [Daddi et al., 2004; Daddi et al., 2005]. Star formation rate (SFR) of galaxies is a challenging area within astronomy due to dust absorption and re-radiation which increases the wavelength (reddening). Nevertheless, it is an important task for accurately observing the processes and mechanisms involved in galaxy formation and evolution.

As a result of data preparation, the constructed data sets have a class attribute representing the discretized value of redshift, and continuous real valued predictor attributes, these were defined as follows:

3 classes: C_1 ($z < 0.8$), C_2 ($0.8 \leq z < 2.2$), C_3 ($z \geq 2.2$)

Colour attributes: B-V, V-R, R-I, I-Z. We shall refer to these as BVmag, VRmag, RImag, IZmag respectively.

BzK atrributes: (B-V)-(V-R), (V-R)-(R-I), (R-I)-(I-Z)

The discretized class values are the same for all datasets. The discretization was performed considering the Lyman Break Method [Steidel, Hamilton, 1992; Steidel et al, 1995], based upon broad spectral features that can be easily identified. These include the Lyman Alpha Break and the Lyman Limits. We observe features at wavelengths of 912Å, 1216Å and 4000Å. These are then subtracted from the central regions in the photometric magnitude regions used in the Hubble surveys. These can be used with photometry to help identify photometric redshift.

The datasets shall be known as Dataset 1 and Dataset 2 from this point on and are defined as:

Dataset 1: [BVmag, VRmag, RImag, IZmag, Class]

Dataset 2: [BVmag, VRmag, RImag, IZmag, BVmag-VRmag, VRmag-RImag, RImag-IZmag, Class]

4. A Genetic Algorithm for estimating redshifts

4.1 Individual Representation

Each individual in the proposed genetic algorithm represents a single classification rule. A rule has the form IF (conditions) THEN (class), a commonly used construct in data mining. The antecedent (IF part) is a conjunction of conditions, where each condition refers to lower and upper bounds for a predictor attribute.

Recall, in section 3, that the predictor attributes given to the system are colours (from the difference in magnitudes B-V, V-R, R-I and I-Z).

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 the name of one of the attributes of the data being mined, in particular a colour attribute in Dataset 1 and a colour or BzK attribute in Dataset 2. In the current version of the GA 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. The rule condition encoded by a gene is satisfied by an example if and only if the value of the corresponding attribute for that example is between the LowerValue and UpperValue thresholds. Finally, Active Flag switches on or off this condition of the rule. The operators are fixed in this version of the

algorithm but are included into the individual encoding as we intend to explore other values for a future version of the algorithm. The individual encoding contains one gene for each attribute of the data being mined. However, only the genes with Active Flag set to 1 ("on") are actually present in the rule decoded from the individual. An example of an individual encoding for the target dataset is as follows (showing one gene per line to facilitate interpretation):

```
0.1200 ≤ BVmag ≤ 0.3000 [1]
0.1100 ≤ VRmag ≤ 0.2800 [0]
0.1800 ≤ RImag ≤ 0.3200 [0]
0.0700 ≤ IZmag ≤ 0.3800 [1]
```

In this example individual, only the conditions referring to the BVmag and IZmag attributes would be present in the decoded rule, so that the decoded rule antecedent would be:

IF 0.1200 ≤ BVmag ≤ 0.3000 AND 0.0700 ≤ IZmag ≤ 0.3800

The decoded conditions are connected by an AND operator to produce an antecedent consisting of a conjunction of conditions.

Note that the class predicted by the rule (in its THEN part) does not need to be encoded in the genome of the individual, since it is fixed throughout the run of the GA, as will be explained in section 4.2.

The values of the genes for each individual are randomly generated when creating the initial population. In that initialisation phase, as well as during the entire run of the GA, the values of the gene elements LowerValue and UpperValue are subject to the restriction that they cannot be smaller than / greater than the minimum / maximum value observed for the corresponding attribute in the training set. This restriction ensures that all rules generated by the GA are sensible and useful to be applied to the test data.

4.2 Sequential Covering

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. Sequential Covering is a technique often used in machine learning and data mining to discover all of the rules required to cover the training example set [Witten, Frank, 2000]. Each value in the set of discretized classes to be predicted is taken in turn as the positive class, the rest collectively become the negative class. The GA then iteratively discovers a rule covering as many positive examples as possible, then removes these examples before repeating the process until all positive examples are covered.

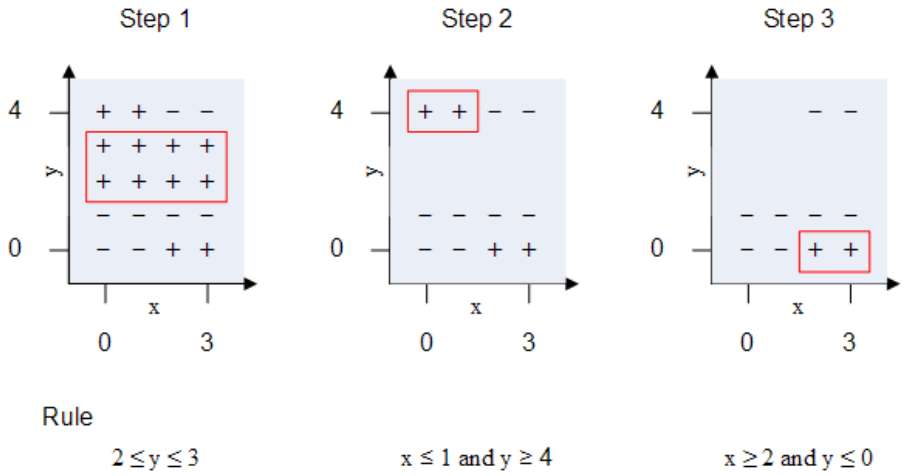

Figure 2: Sequential Covering. This shows 3 steps in the process, where the best positive class is identified from the population of rules and the positive examples covered by that rule are then removed from the data and the GA is run again. This process iteratively builds up a set of rules covering all of the positive examples (which identify just one class to be predicted).

Once all of the positive examples have been classified, and a set of rules has been created, the entire process is then repeated using the next class value. This iterative process is summarized in Figure 2.

Once a set of rules exists covering all of the training examples, they are ordered by the classification accuracy of those rules (i.e. their fitness). In order to classify examples that are not covered by any rule, the system enables a last "default rule", predicting the class covering most training examples.

When using these rules on the test set, for each test example, all rules covering that example are initially 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 (i.e. the highest rule in the ordered set of rules). Finally, if no rules cover the example, the default rule is chosen, predicting the class which has the most examples in the training set.

4.3 Fitness Function

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 (also known in the astronomical community as reliability) and completeness (or true positive rate) are calculated as shown below, the product of these derives the fitness for each rule.

```
Confidence = TP/(TP+FP)
Completeness = TP/(TP+FN)
```

```
Fitness = Confidence * Completeness
```

where TP, FP, TN and FN refer to the following entries of the confusion matrix: True Positive, False Positive, True Negative and False Negative.

Table 1 shows a confusion matrix for a rule of the form IF A THEN C, each quadrant of the matrix has the following definition [Freitas, 2002]:

TP (True Positives)	Number of instances satisfying A and belonging to class C
FP (False Positives)	Number of instances satisfying A but not belonging to class C
FN (False Negatives)	Number of instances not satisfying A but belonging to class C
TN (True Negatives)	Number of instances not satisfying A nor belonging to class C

Table 1: Confusion Matrix

		Actual Class	
		C	*not C*
Predicted Class	*C*	TP	FP
	not C	FN	TN

4.4 Genetic Operators (Crossover and Mutation)

4.4.1 Crossover

The crossover operator is used to evolve the current generation of individuals into a new set of individuals with parents selected from the current population. The crossover method used in our GA is uniform crossover. This works by taking each gene (or rule condition) in turn and randomly choosing the value of that gene (the sextuplet) in one of the parents for one child and the value of that gene in the other parent for the other child. Therefore the two children will be made up from parts of the two parents. The crossover probability used here was 70%.

Uniform crossover was chosen as there is no positional bias as can be experienced using n-point crossover methods [Falkenauer, 1999; Syswerda, 1989]. In the application domain of the proposed GA, as well as in the classification task of data mining in general, the set of attributes has to be treated as a set in the mathematical sense of the term, where there is no ordering between the attributes. Therefore, the lack of positional bias associated with uniform crossover is appropriate for this application domain.

4.4.2 Mutation

Mutation in a GA maintains diversity in the candidate solutions and therefore allows greater exploration of the data space. In our GA, each individual's gene elements probabilistically undergo mutation. The elements of each gene 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 the attribute values by up to 0.5, and to the active flag by flipping it to set whether this particular gene will be active, and therefore have that attribute included in the rule. Hence, the mutation of the active flag can effectively add or remove any condition within the current candidate rule, contributing to maintain a set of candidate rules with diverse lengths in the population.

4.5 Selection Method

Selection is the method used to pick parents that will undergo genetic operators such as crossover and mutation, in order to generate offspring in the next generation. In the proposed GA tournament selection is used, where five individuals are chosen at random and then pitted against each other in a virtual tournament. The best (highest fitness) survives and goes through for mating with another individual winner from another tournament.

5. Computational Results

In order to evaluate the proposed GA, we compared its performance against the performance of C5.0 [RuleQuest, 2005]. C5.0 is an industrial strength state of the art commercial decision tree and rule induction product from RuleQuest Research, developed by Ross Quinlan as the successor to his very successful and widely used ID3 and C4.5 systems. We used the C5.0 implementation available as a built in resource within the well-known Clementine data mining tool.

C5.0 can generate a classification model expressed either as a decision tree or as a set of rules. In order 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 [Witten, Frank, 2000], as usual in machine learning and data mining. We used the default parameters of C5.0 and of the GA – i.e., we did not try to optimise the parameters 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 class values used were the same for both datasets and were based upon spectral features as described in section 3. The probabilities of application of genetic operators were as mentioned in section 4. 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 average classification accuracies across the 100 runs of the GA and the 10 runs of C5.0 are shown in Table 2:

Table 2: Classification accuracies for the runs of the GA and C5.0

Algorithm	Dataset 1	Dataset 2
GA	(93.16 +/- 0.46)%	(90.4 +/- 3.04)%
C5.0	(90.73 +/- 0.63)%	(89.852 +/- 2.23)%

The results upon Dataset 1 have >99% confidence that they are statistically significantly different and not by chance, as measured by a Student t-test. However, the difference in the accuracies of the GA and C5 in Dataset 2 are not statistically significant as measured by a Student t-test.

A measure of simplicity was taken of the rules, this is the number of discovered rules and the total number of terms, or conditions, for all discovered rules, as usual in the data mining literature. 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. The full simplicity results are shown in Table 3:

Table 3: Simplicity of rules generated by the GA and C5.0

Algorithm	Dataset 1		Dataset 2	
	Rules	Terms	Rules	Terms
GA	8.43 (+/- 1.64)	10.01 (+/- 2.66)	8.8 (+/- 1.66)	15.1 (+/- 4.39)
C5.0	16.5 (+/- 2.58)	41.8 (+/- 5.69)	19.1 (+/- 3.73)	57.9 (+/- 15.61)

The measurements of the number of rules and total number of terms in all rules upon Dataset 1 and Dataset 2 have >99% confidence that they are statistically significantly different and not by chance, as measured by a Student t-test.

Work is ongoing to interpret the classification rules discovered by the GA in the context of prior astronomical knowledge, such as the passage of known redshifted spectral features through the filter passbands.

6. Conclusion and Future Research

The first results of the new GA compared to C5.0 show an improvement in the accuracy with the GA performing with an accuracy over 2.4 percentage points higher than C5.0 for Dataset 1, a statistically significant difference. 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.

Results using BzK conditions showed a decrease in accuracy for both the GA and C5.0. This was surprising because of the relationship between redshift and SFR in galaxies shown in [Daddi et al., 2004]. In our future work we will continue BzK tests when we have additional data from other surveys.

The only attributes used in these runs were the constructed colour and BzK 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), which will extend upon our use of BzK relationships. Relational conditions are used in astronomy, for example in finding star forming galaxies [Daddi et al., 2004]. A third research direction is to perform template matching using HYPER-Z [Bolnzella et al. 2000] 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

[Appenzeller, 2005] Appenzeller, I. FORS consortium; The FORS Deep Field (FDF) [http://www.lsw.uni-heidelberg.de/users/jheidt/fdf/fdf.html], Visited October 2005

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

[Bowman et al., 1993] Bowman, B. et al. Reasoning about naming systems. ACM Trans. Program. Lang. Syst., 15, 5 (Nov. 1993), 795-825, 1993.

[Collister, Lahav, 2003] Collister, A. A., Lahav, O. ANNz: estimating photometric redshifts using artificial neural networks, astro-ph, 0311058, 2003

[Connolly et al., 1995] Connolly, A.J. et al. Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry, astro-ph/9508100, 1995

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

[D'Odorico et al., 2005] D'Odorico et al. European Space Agency; The NTT SUSI Deep Field [http://www.eso.org/science/ndf/], Visited October 2005

[Daddi et al., 2004] Daddi et al., The Population of BzK Selected ULIRGs at z ~ 2, astro-ph/0507504v1, 2005

[Daddi et al., 2005] Daddi et al., Star-forming and Passive Galaxies, ApJ, 617, 746, 2004

[Ding, Marchionini, 1997] Ding, W., Marchionini, G. A Study on Video Browsing Strategies. Technical Report UMIACS-TR-97-40, University of Maryland, College Park, MD, 1997.

[ESO, 1998] ESO Education & Public Relations Department. Deep Galaxy Counts and Photometric Redshifts in the HDF-S NIC3 Field [http://www.eso.org/outreach/press-rel/pr-1998/pr-20-98.html], 1998, Visited on 15 August 2004

[ESO, 1999] ESO Education & Public Relations Department. The FORS/ISAAC Cluster Deep Field [http://www.eso.org/outreach/press-rel/pr-1999/phot-09-99.html], 1999, Visited on 15 August 2004

[Falkenauer, 1999] Falkenauer E., The Worth of the Uniform, CEC-99, 1999

[Fayyad, 1996] Fayyad U. M., Chapter 19: Automating the Analysis and Cataloging of Sky Surveys, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996

[Ferguson, 2005] Ferguson H., Space Telescope Science Institute. The Hubble Deep Field [http://www.stsci.edu/ftp/science/hdf/hdf.html], Visited October 2005

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

[Freitas, 2002] Freitas A. Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, 2002

[Giavalisco et al., 1997] Giavalisco M. et al. The Hubble Deep Field: Number Counts, Color-Magnitude and Color-Color Diagrams, 1997

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

[Kwedlo, Kretowski, 2000] Kwedlo W., Kretowski M. An Evolutionary Algorithm Using Multivariate Discretization for Decision Rule Induction, Proceedings of Evolutionary Computations on Principles of Data Mining and Knowledge Discovery (PKDD'00), Springer LNCS 1910, 2000

[Kwedlo, Kretowski, 1998] Kwedlo W., Kretowski M. Discovery of decision rules from databases: an evolutionary approach. Principles of Data Mining and Knowledge Discovery, PKDD'98. Nantes, France. Springer LNCS 1510, 1998

[Lahav et al., 1996] Lahav O. et al. Neural Computation as a tool for galaxy classification: methods and examples, MNRAS, 283, 207L, 1996

[Peacock, 1999] Peacock J. A. Cosmological Physics, Cambridge University Press, 1999

[Rengelink, 1998] Rengelink R. AXAF Field: Deep Optical-Infrared Observations, Data Reduction and Photometry, ESO Imaging Survey, 1998

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

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

[Stanway et al., 2003] Stanway E. R. et al. Lyman Break Galaxies and the Star Formation Rate of the Universe at z ~ 6, 2003

[Staneck, 1999] Stanek, R., Photometry [http://astrwww.astr.cwru.edu/nassau/reference/photometry.html], CWRU Astronomy Dept, 1999, Visited on October 2005

[Steidel, Hamilton, 1992] Steidel C. C., Hamilton D. Deep imaging of high redshift QSO fields below the Lyman limit. I - The field of Q0000-263 and galaxies at Z = 3.4, AJ, 104, 941-949, 1992

[Steidel et al, 1995] Steidel C. C. et al. Lyman Imaging of High-Redshift Galaxies.III.New Observations of Four QSO Fields, AJ, 110, 2519, 1995

[Steidel et al., 1996] Steidel C. C. et al. Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at redshifts z>3, AJ, 462, L17–L21, 1996

[Syswerda, 1989] Syswerda G., Uniform Crossover in Genetic Algorithms, ICGA-89, 1989

[Wang et al., 1998] Wang et al. A Catalog of Color-based Redshift Estimates for Z <~ 4 Galaxies in the Hubble Deep Field, AJ 116, 2081, 1998

[Williams et al., 1997] Williams R. et al. The Hubble Deep Field: Images, 1997

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

[Witten, Frank, 2000] Witten I. H., Frank E. Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2000