Astronomical Data Analysis Software and Systems XVIII
ASP Conference Series, Vol. 411, © 2009
D. Bohlender, D. Durand and P. Dowler, eds.

Star-Galaxy Classification Using Data Mining Techniques with Considerations for Unbalanced Datasets

Peter J. O'Keefe, Michael G. Gowanlock, Sabine M. McConnell,[1] and David R. Patton[2]

Abstract. We used a range of data-mining techniques in an effort to improve the classification of stars and galaxies for imaging data from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS), and extracted with SExtractor. We found that the Artificial Neural Network (ANN) achieved higher accuracies than Support Vector Machines, but was outperformed by the Random Forest and Decision Tree data-mining techniques on 5000 randomly sampled objects. This has potentially negative implications for SExtractor which uses an ANN to produce a measure of stellarity for each object. We found that the classification of stars and galaxies can be improved by voting (between Decision Trees, Random Forests and ANNs) and using balanced datasets. For the balanced datasets that we created, the three data mining techniques agreed over 80% of the time on the type of object.

1. Introduction

Classification of stars and galaxies from astronomical images is a tedious, yet essential process for researchers in the field of astronomy. Due to the subjectivity inherent in manual classification, the outcome of the process is highly dependent on the individual analyzing the image. The quality of the image and other systematic factors of various astronomical surveys must also be taken into consideration in order to correctly classify an object. Advanced image processing techniques and powerful learning algorithms make automated star and galaxy classification a faster alternative to its manual counterpart (Philip et al. 2002).

There are a variety of tools that can assist or fully automate the classification of stars and galaxies from astronomical images. SExtractor (Bertin & Arnouts 1996) uses an Artificial Neural Network (ANN) to perform star-galaxy classification. In the default configuration, nine attributes (eight isophotal attributes and one attribute related to peak intensity) are used to classify the objects. One of the outputs is a parameter known as "stellarity". Objects with a value for this parameter closer to 0 are more likely to be a galaxy and objects closer to 1 are more likely to be a star. Tools such as SExtractor can produce accurate results but when conditions are not ideal, the results produced are often very poor and in some cases completely unreliable. Depending on the quality of the image, SExtractor may not be able to clearly distinguish between stars and galaxies, resulting in the majority of classified objects remaining in the un-

[1]Department of Computing & Information Systems, Trent University, ON, Canada

[2]Department of Physics & Astronomy, Trent University, ON, Canada

known region (when the mean stellarity is approximately 0.5). The motivation of this research is to use a range of advanced data-mining methods in an effort to improve the classification of objects returned from SExtractor.

2. The Data

For this project we used data from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS). The specific image we used was the i-band image for the D1 field together with the corresponding weight image. SExtractor produced approximately 590,000 objects from this image. To make this dataset more manageable, 5,000 objects were randomly sampled from the entire dataset.

After analyzing the ranges of stellarity for the objects within the unknown group in more detail, the results still appeared to be somewhat misleading. First, approximately 290,000 objects fell into the stellarity range above 0.5 and below 0.9. According to SExtractor, these objects are more likely to be a star. The second range contained objects with a stellarity between 0.1 and 0.5, with approximately 125,000 objects in this range. The majority of objects were closest to 0.5 but a large number of objects were misclassified as more likely a star rather than a galaxy. Examination of the image shows that this cannot be the case.

About 30% of the original dataset produces reliable results and of this 30%, 75% were stars and 25% were galaxies. This 30% of what we considered to be "reliable" data was used as training data to verify the accuracy of the different models created. The difference in the number of objects between these two classes is not only large, but it is also misrepresentative of the entire image. Investigating the problem by taking into consideration class imbalance will enable us to build an accurate model.

3. Class Imbalance

Class imbalance is a situation where one class in the dataset is represented by a large number of objects and the other class is represented by very few objects. In class imbalance problems the minority class will typically be under-represented and therefore misclassified. Creating balanced datasets is one method which has been successful in remedying the imbalance problem (Japkowicz 2000). In training sets, we therefore created three datasets using sampling techniques as follows. The first dataset (DS1) consisted of all of the galaxies and an undersample of the stars (383 galaxies, 383 stars). The second dataset (DS2) consisted of all of the stars and an oversample of the galaxies (1128 stars, 1128 galaxies). The third dataset consisted of half of the stars and the same number of galaxies, which had to be oversampled (546 stars, 546 galaxies).

4. Method

We used SExtractor to produce a measure of stellarity for each object in a portion of the CFHTLS data. The stellarity parameter produced will give a star a value close to 1 and a galaxy a value close to 0. We assumed that a stellarity value between 0-0.1 was a galaxy and a stellarity value between 0.9-

1.0 was a star. The class distributions were as follows: 70.5% were neither a star nor galaxy, 7.5% were galaxies and 21.9% were stars. Although SExtractor uses an ANN to classify the objects, we were interested in voting between the Random Forest (RF), Decision Tree (DT) and ANN data-mining techniques using balanced datasets to see if we could determine which factors were primary in contributing to the classification of stars and galaxies. The Support Vector Machine (SVM) classification technique was tried but produced very poor results which is partly due to the unreliable or noisy data.

We took 5000 randomly sampled objects from the dataset produced by SExtractor and used it to create three models in the Waikato Environment for Knowledge Analysis (WEKA)[3] (Witten & Frank 2005) corresponding to the DT, RF, and ANN data-mining techniques. We ran our balanced datasets as test sets against the three models to extract the predictions used in our comparison. We then randomly selected 5000 objects from the original data each time we created a balanced dataset to better represent the entire sample.

It is our hypothesis that if an object's predicted type (star or galaxy) is voted upon and it matches the actual value as determined by WEKA by comparing it to 5000 randomly sampled objects, then such consistency would result in particular factors that would make star and galaxy classification more accurate. However, if the confusion matrix shows that some of the stars and galaxies in the balanced datasets that contain only known stars and galaxies are incorrectly classified, then star-galaxy classification will be difficult for this particular dataset.

5. Results

The four data-mining techniques performed relatively well on a dataset of randomly sampled objects from the CFHTLS data. We used this dataset to produce four models corresponding to each data-mining technique. Considering that SExtractor uses an ANN to classify objects, it is interesting that the RF and DT methods outperformed the ANN in WEKA. We assumed that the data would be inherently biased in favour of the ANN.

It appears as though there is not a particular way to balance the data in this dataset to achieve the best results. The three techniques agreed that an object was a star or a galaxy approximately 82% of the time. Also, the average deviation of the model accuracy between the three balanced datasets was low, reinforcing the notion that the three balanced datasets all performed relatively equally (see Table 1). The datasets contained a few objects for which the three different data-mining techniques could not agree on a type, suggesting these instances should be treated as outliers and possibly removed from the datasets.

6. Conclusion

We found that voting between multiple data-mining techniques in SExtractor can improve the accuracy of star-galaxy classification. Considering that the

[3]http://www.cs.waikato.ac.nz/ml/weka/

Table 1.: Accuracy of the techniques when tested against balanced datasets.

	DS1 [%]	DS2 [%]	DS3 [%]	Average Deviation [%]
DT	80.03	80.54	80.68	0.26
RF	82.90	82.98	82.97	0.03
NN	72.58	72.52	75.82	1.45
SVM	36.81	36.44	37.82	0.53

ANN was outperformed by the DT and RF data-mining techniques, allowing for more data-mining methods to be utilized in SExtractor can be beneficial to the astronomical community. The principle data investigated did not benefit from any particular balancing scheme, however the balanced data did refine our interpretations of the image, and led us to conclude that balancing options in SExtractor would improve the automated classification process for some astronomical images. In the future, we would like to consider adding more data mining classification techniques to the voting process. Perhaps a voting scheme could be implemented that gives a greater weight to the technique(s) that perform the best on an initial sample of the data. In practice, this would mean that the technique that most accurately fits the data would have a greater weight in the classification process.

Acknowledgments. This work was funded in part by NSERC of Canada via a Discovery Grant to DRP. Based on observations obtained with MegaPrime/ MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Science de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, a collaborative project of NRC and CNRS.

References

Bertin, E. & Arnouts S. 1996, A&AS, 117, 393
Japkowicz, N. 2000, ICAI, 1, 111
Philip, N. S., Wadadekar, Y., Kembhavi, A. & Joseph, K. B. 2002, A&A, 385, 1119
Witten, H. I. & Frank, E. 2005, Data Mining: Practical machine learning tools and techniques