Automated Classification of Astronomical Objects in Multispectral Wide-Field Images

Jorge de la Calleja *
Computer Science Department
I.N.A.O.E.
Tonantzintla, Puebla, 72840, Mexico

Olac Fuentes
Computer Science Department
University of Texas at El Paso
El Paso, Texas, 79968, U.S.A.

Abstract

In this paper we present an automated method for classifying astronomical objects in multispectral wide-field images. The method is divided into three main tasks. The first one consists of locating and matching the objects in the multispectral images. In the second task we create a new representation for each astronomical object using its multispectral images, and also we find a set of features using principal component analysis. In the last task we classify the astronomical objects using neural networks, locally weighted linear regression and random forest. Preliminary results show that the method obtains over 93% accuracy classifying stars and galaxies.

Introduction

Astronomical surveys in different multispectral bands, such as the Sloan Digital Sky Survey (SDSS)[1], produce enormous amounts of data that require automated tools for any analysis. One of the problems in astronomical data analysis is the classification of astronomical objects. An astronomical wide-field image normally contains from tens to thousands of objects such as stars, galaxies, and nebulas, among others. Multispectral imaging refers to acquiring several images of the same scene using different spectral bands. Analyzing wide-field images has been and still is of great importance in astrophysics: from studies of the structure and dynamics of our galaxy, to galaxy formation and evolution, to the large scale structure of the Universe (Andreon *et al.* 2000).

There has been much research work using machine learning algorithms in order to classify astronomical objects, including galaxies (Bazell & Aha 2001; De la Calleja & Fuentes 2004; Lahav 1996), stars (Bailer-Jones, Irwin, & von Hippel 1998), star/galaxy discrimination (Andreon *et al.* 1999; Mahonen & Hakala 1995), and binary stars (Weaver 2000). Recently, Zhang and Zhao used learning vector quantization, support vector machines and single-layer perceptrons to classify AGNs, stars and galaxies, using data from optical, X-ray and infrared band (Zhang & Zhao 2004).

*The authors want to thank CONACyT for partially supporting this work under grant 166596

[1]http://www.sdss.org

We propose a fully automated method to classify astronomical objects in multispectral wide-field images. This method first locates and matches the objects in the different multispectral images. Then, it creates a new representation for each object using its multispectral images. After that, it employs principal component analysis to reduce the dimensionality of the data and to find relevant information. Finally, it uses machine learning algorithms to classify the objects.

The paper is organized as follows: in Section 2 we describe the classification method. Section 3 presents experimental results. Some conclusions and directions of future research are presented in Section 4.

The method

First, for each multispectral image we separate the objects from background applying a threshold. Then, we locate the astronomical objects using the coordinates (x_1, y_1, x_2, y_2) of the bounding rectangles that enclose them. We assume that all the pixels that appear different from the background correspond to astronomical objects. Thus, our algorithm (based on the *flood-fill* algorithm (Hearn & Baker 1997)) proceeds as follows: We examine every location in the image, then when we find a pixel that is different from the background, we search its neighbors in the right, left, up, and down directions, and mark these pixels. We store the positions of these neighbors to examine again their neighbors. When we find a left neighbor and its position x_1 is smaller than the previous one, we will change it. We do the same for y_1, x_2 and y_2, but considering up, right, and down, respectively. This process is repeated until the image is fully examined.

The main idea of our algorithm to match objects is the following: We search for the largest object in the multispectral images. Then, we use its coordinates (x_1, y_1, x_2, y_2) to find the same object, but in the other images. Almost always the object will not appear exactly in the same position, then we use the Euclidean distance to match the closest object to our interest point. Sometimes we may not match any object because objects don't emit energy at all wavelengths, thus resulting in black regions in some of the images; in these cases we assign as match point the location of the largest object. This process is repeated until all the objects have been matched.

Table 1: Accuracies of LWLR, ANNs and RF for the data sets.

	DB1			DB2		
PCs	LWLR	ANN	RF	LWLR	ANN	RF
1	81.90	**89.57**	79.70	76.65	83.20	77.50
2	87.04	87.04	83.76	87.91	83.47	84.73
3	**91.28**	88.14	**87.57**	**93.71**	86.37	**87.92**
6	90.56	84.09	84.05	90.62	**87.59**	87.33

We create a new representation of each object as follows: We build a vector for each object using its multispectral images. Then, these vectors are concatenated to create a matrix of size of $m \times n \times r$, i.e. the size of the vector by the number of multispectral representations. After that, we use principal component analysis (PCA) to reduce the dimensionality and to generate features.

Finally, we use artificial neural networks (ANNs), locally weighted linear regression (LWLR) and random forest (RF) to classify the astronomical objects. Each method takes as input parameters the projection of the new representation of the objects onto set of a few principal components. Details about LWLR and RF can be found in (De la Calleja & Fuentes 2004) and (Breiman 2001), respectively.

Experimental Results

We tested our method using images taken in five wavelengths: one in ultraviolet, one using a green filter, one using a red filter, and two using infrared filters. We used two data sets obtained from the SDSS, the first one (DB1) contained 62 stars and 7 galaxies, and the the second one (DB2) had 141 stars and 28 galaxies. We used different numbers of principal components (PCs), considering the information that they represent. Thus, we used 1, 2, 3, and 6 PCs, that represent about 75%, 85%, 90% and 95% of the original information in the data sets.

We implemented locally weighted linear regression in Matlab, and used the feedforward network that is implemented in the Matlab Neural Network Toolbox. Also we use the random forest classifier that is implemented in WEKA [2].

The accuracies that we show in this section were obtained by averaging the results of 5 runs of 10-fold cross validation for each machine learning method. Table 1 shows the accuracies obtained by LWLR, ANNs, and RF considering both data sets. We can see that LWLR obtained the best accuracies in four cases, while ANNs obtained the second best results. Also, we can notice that using three PCs we obtain the best results. Finally, in Table 2 we show confusion matrix of the best single results for LWLR.

Conclusions and Future Work

We have presented a method for classifying astronomical objects in multispectral wide-field images in a fully automated manner. Our results are comparable with the best ones reported in the literature. Locally weighted linear regression obtained the best results. We also conclude that a small set of principal components is enough to classify the astronomical objects. Future work includes testing the method for classifying more types of astronomical objects and dealing with unbalanced data sets.

[2]WEKA is a software package that can be found at www.cs.waikato.ac.nz/ml/weka

Table 2: Confusion matrix obtained by LWLR for the data sets.

	Galaxies	Stars	Galaxies	Stars
Galaxies	3	4	23	5
Stars	1	61	2	139

References

Andreon, S.; Gargiulo, G.; Longo, G.; Tagliaferri, R.; and Capuano, N. 1999. Neural nets and star/galaxy separation in wide field astronomical images. In *Proceedings of IJCNN99*.

Andreon, S.; Gargiulo, G.; Longo, G.; Tagliaferri, R.; and Capuano, N. 2000. Wide field imaging - I: Applications of neural networks to object detection and star/galaxy classification. *MNRAS* 319:700–716.

Bailer-Jones, C.; Irwin, M.; and von Hippel, T. 1998. Automated classification of stellar spectra. ii: Two-dimensional classification with neural networks and principal components analysis. *MNRAS*.

Bazell, D., and Aha, D. 2001. Ensembles of classifiers for morphological galaxy classification. *ApJ* 548:219–233.

Breiman, L. 2001. Random forests. *ML* 45(1):5–32.

De la Calleja, J., and Fuentes, O. 2004. Machine learning and image analysis for morphological galaxy classification. *MNRAS* 349:87–93.

Hearn, and Baker. 1997. *Computer Graphics*. Prentice Hall.

Lahav, O. 1996. Artificial neural networks as a tool for galaxy classification. In *Proceedings in Data Analysis in Astronomy*.

Mahonen, P., and Hakala, P. 1995. Automated source classification using a kohonen network. *ApJ* 452:L77–L80.

Weaver, B. 2000. Spectral classification of unresolved binary stars with artificial neural networks. *ApJ* 541:298–305.

Zhang, Y., and Zhao, Y. 2004. Automated clustering algorithms for classification of astronomical objects. *ApJ* 422:1113–1121.