Report

Problem statement:

Understanding how and why we are here is one of the radical mysteries for the human race. Fragment of the answer to this problem lies in the origins of galaxies, such as our own Milky Way. As per reports from The University of Chicago, the Sloan Digital survey conducted by Apache Point Observatory will produce more than 50 million images of galaxies in the near future. Stratification of these images is usually done by visual examination of photographic plates. Interpreting the distribution, location and types of galaxies as a function of shape, size, and colour are critical pieces for evaluating our place in the universe. We plan to build an algorithm to classify properties of galaxies with a degree of probability close to the annotation done by volunteers as part of a classification project.

Literature Review:

Image classification has been a widely studied area in the field of machine learning. However, most of the image datasets and machine learning algorithms are targeted at images of natural scenes and objects, which is different from galaxy images we are focussing here. In [1] two computations schemes have been proposed for Hubble galaxy classification. The first scheme uses geometric shape as the feature for classification, and the other schemes use the image pixel values as features and artificial neural network for classification.

[2] used a neural network and locally weighted regression method as the weak learners and implemented bagging as the ensemble method for classification. In [3], three different techniques have been used as classification which are Naive Bayes, rule-induction algorithm C4.5, and Random Forest. Among the three implementation schemes, Random Forest produced the best result.

Dataset Details:

The Galaxy Zoo Dataset used is taken from the Kaggle challenge. The dataset consists of 61,578 RGB galaxy images of 424 x 424 pixels. Few sample images from the dataset are shown in Fig. 1. Each of these images has been annotated by human volunteers and accordingly probabilities of corresponding 37 categories are provided.

These 37 categories correspond to 11 morphological questions such as "How many spiral arms does the galaxy have?". Each answer will have a probability associated with it calculated with the help of human annotation. Furthermore, the dataset will be split into 70% for training and 30% for testing.

Proposed Architecture:

There were 11 different questions with cumulatively 37 answers. We treated each question as a separate prediction problem and proceeded further.

The pipeline of architecture is shown in the figure. We used HOG,LBP and image moments as features. For each of the proposed prediction algorithms

namely, Linear Regression, Lasso Regression, and Ridge Regression we analyzed the effect of each feature on different questions separately.

The analysis is shown in the fig. The analysis against RMSE led to the following feature selection for the proposed architecture as shown in the table.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Regression | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 |
| Linear | lbp | lbp | im | lbp | im | im | lbp | im | im | im | im |
| Lasso | any | any | any | any | any | any | any | any | any | any | any |
| Ridge | hog | hog | hog | hog | hog | im | hog | im | im | im | im |

References :

[1] Goderya, S.N., Lolling, S.M., “Morphological Classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational Scheme.” Astrophysics and Space Science. Vol. 279, no. 4, pp. 377 –387.

[2] Calleja, J., Fuentes, O., “Machine Learning and Image Analysis for Morphological classification of galaxies.” Monthly Notices of the Royal Astronomical Society,Vol. 24, 2004, pp. 87‐93.

[3] Calleja, J., Fuentes, O., “Automated Classification of Galaxy Images”, Lecture Notes in Computer Science, Vol. 3215, 2004, pp. 411‐418.