

Extraction and Classification of Objects from Astronomical Images in the Precense of Labeling BIAS. *Dr. Guillermo Cabrera, DIM-DCC, FCFM, U-Chile.*

ABSTRACT:

Massive amount of astronomical data is starting to flow from the new generation of telescopes. Due to the big distance the astronomical sources are at, though we can have an estimate of their characterization, it is impossible to know their real properties.

In this work we address the problem of extracting light profiles from astronomical objects and use them to classify them. As in many other real world classification problems, we use labels created by human annotators, which are usually assumed to be equally skilled to classify objects regardless the parameter space where these objects fall. However, real data is never perfect, and labels given by these annotators can be systematically biased due to poor quality of the observed data they are labeling. This is particularly true for the case of galaxy morphology classification, were small, faint and distant galaxies are often classified as ellipticals because of poor image resolution. We present two methods for obtaining real labels from biased data: unbiased parameter selection and an E-M approach which jointly learns a classification model, estimates the bias, and obtains new debiased labels. Experiments show that both methods are able to achieve unbiased labels when there is no absolute ground truth.

RESUMEN:

Una enorme cantidad de datos astronómicos está comenzando a fluir de la nueva generación de telescopios. Debido a la gran distancia a las fuentes astronómicas, aunque podemos tener una estimación de su caracterización, es imposible conocer sus propiedades reales.

En este trabajo se aborda el problema de la extracción de perfiles lumínicos de objetos astronómicos y su uso para clasificación. Como en muchos otros problemas de clasificación reales, usamos etiquetas creadas por anotadores humanos, los cuales por lo general se asume que son igualmente hábiles para clasificar objetos independientemente del espacio de parámetros donde estos objetos caen. Sin embargo, los datos reales no son perfectos, y estas etiquetas pueden estar sesgadas sistemáticamente debido a la mala calidad de los objetos observados. Esto es particularmente cierto para el caso de la clasificación morfológica de galaxias, donde galaxias pequeñas, débiles y lejanas se clasifican a menudo como elípticas debido a la mala resolución de la imagen. Presentamos dos métodos para obtener etiquetas reales a partir de datos sesgados: selección de parámetros no sesgados y un enfoque Expectation-Maximization que aprende conjuntamente un modelo de clasificación, estima el sesgo, y obtiene nuevas etiquetas no-sesgadas. Los experimentos muestran que ambos métodos son capaces de lograr etiquetas no-sesgadas cuando no existe una verdad absoluta.