Searching for quasars in AllWISE data

Artem Poliszczuk
National Centre for Nuclear Research

4th Cosmology School, Cracow, 2018

Unified AGN model

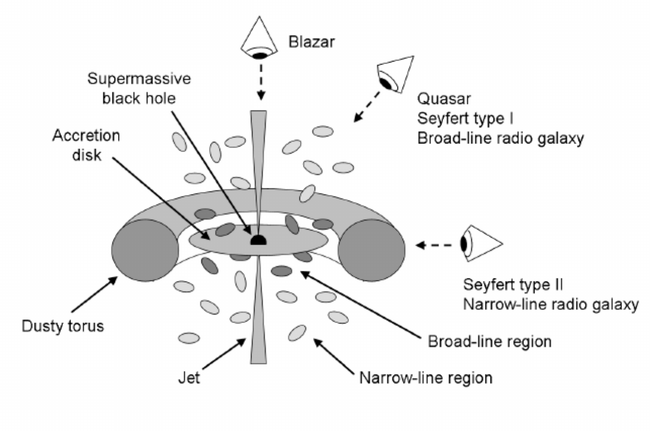

Figure: Unified AGN model. Credit: Zackrisson et al. 2005

Support Vector Machines (SVM)

Support Vector Machines classification algorithm (V. Vapnik 1995)

- supervised learning algorithm: need to give an example input with known labels. Tries to learn a rule that maps input to the labels
- higher performance then traditional learning algorithms
- powerful tool for solving classification problems

The SVM method

We are given a set **S** fo labeled training points:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_k, y_k) \tag{1}$$

Each **training point** $\mathbf{x}_i \in \mathbb{R}^N$ belongs to either two classes and is given a **label** y $\in \pm 1$ for i $= \overline{1, k}$

Problem

In most cases we can't find a suitable hyperplane in an input space

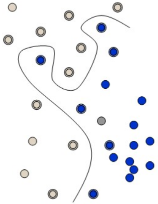

Figure: Credit: www.dtreg.com

Mapping to a higher dimensions

Solution

Mapping the input space into a higher dimension feature space and searching the optimal hyperplane

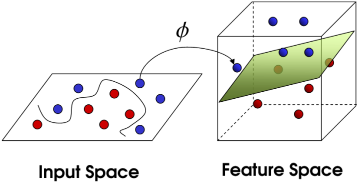

Figure: Credit: www.dtreg.com

$$\phi : \mathbb{R}^N \longrightarrow Z \tag{2}$$

Example: 1D binary classification

Finding the optimal hyperplane

- For the lieary separable set - unique opitmal hyperplane with maximized margin
- Solution of the optimal hyperplane can be written as acombination of a few input points that are called **support vectors**
- New data points class assigned based on their position relative to the boundary

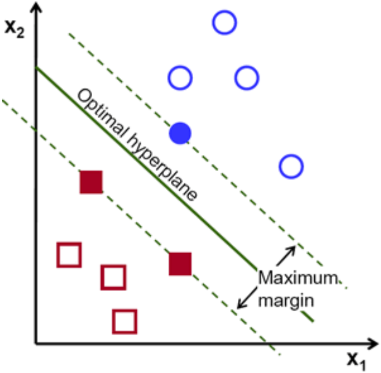

Figure: Credit: docs.opencv.org

AllWISE data

Wide-field Infrared Survey Explorer (WISE). NASA IR Satellite (launched in 2009) All Sky survey in four passbands:

- 3.3 μm (W1)
- 4.7 μm (W2)
- 12 μm (W3)
- 23 μm (W4)

AllWISE Catalog:
747 million objects.

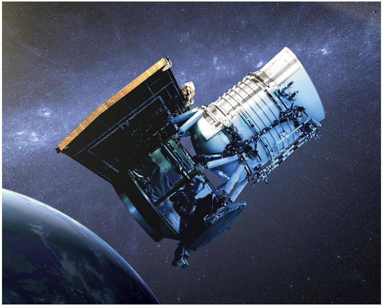

Figure: Credit:www.nasa.gov

AllWISE-SDSS cross-match

- In order to obtain labeled data set one has to cross-match AllWISE catalog.
- Due to need for high statistics: SDSS DR14.
- Around 3 million objects (380 000 QSO).
- Selection effect.

Input parameter space

Parameters used in training: Kurcz et al. 2016.

- W1, W2
- Concentration = w1mag1-w1mag3

w1mag1 - 5.5" radius aperture magnitude
w1mag3 - 11.0" radius aperture magnitude
Binary classification: QSO(5k) vs. Rest(5k=2.5k stars + 2.5k galaxies)

Problematic distribution of final catalog

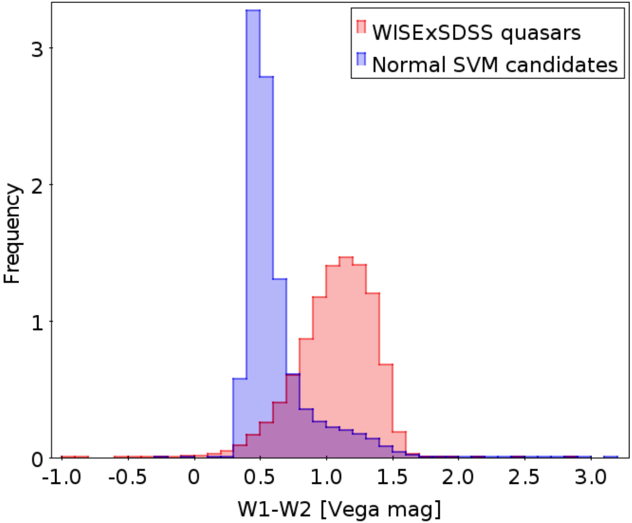

Figure: Generalization on AllWISE Data

Classification of validation set

Classification of AllWISExSDSS14 not used in the training.
Completness: 94%, purity: 83%

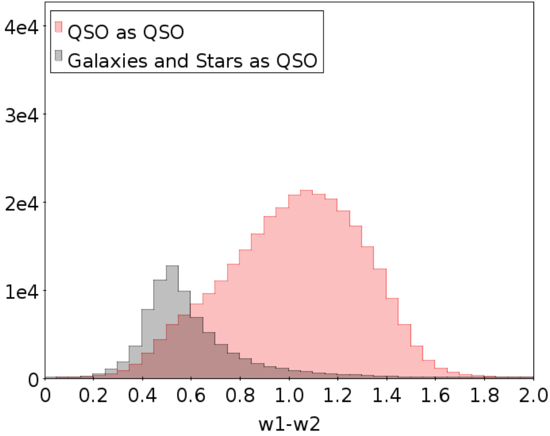

Figure: Validation set

Using probability as additional feature

Probability based on the distance from decision surface can be used as additional feature in the secondary classification.

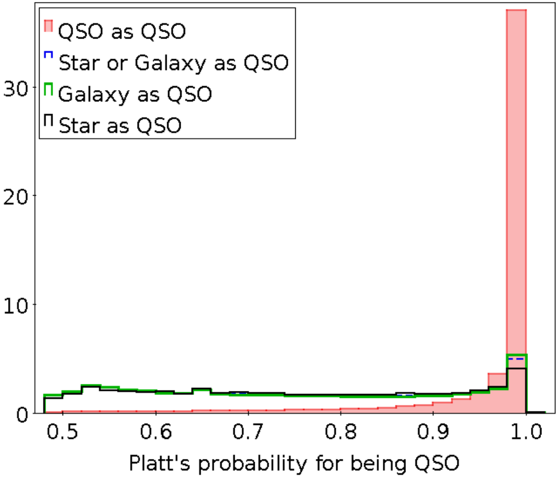

Figure: Posteriori Platts probability of being QSO

Using probability as additional feature. Results

Completness: 94% → 80%, purity: 83% → 97%

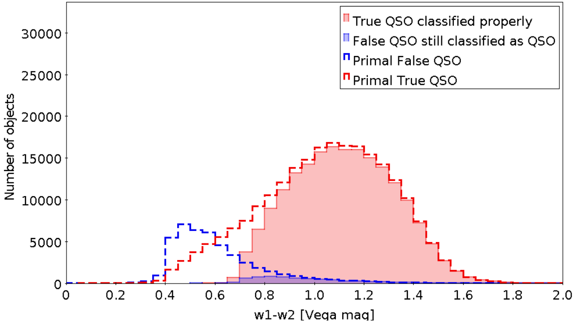

Figure: Second iteration with added probabilities

Summary

- Understanding the distribution problem.
- Satisfactory beginning results.
- A lot of things to test and improve.