

Spectral Classification of Quasar Subject to redshift: A Statistical Study

Author: Prithwish Ghosh and Shinjon Chakraborty
Presenter: Shinjon Chakraborty

Department of Statistics
ghosh.prithwish1999@gmail.com
shinjonchakraborty07@gmail.com

Table of Contents

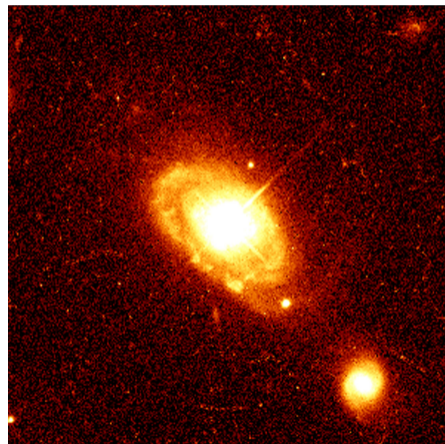
- 1 Introduction & Motivation
- 2 Theory
- 3 Results & Discussion
- 4 Conclusion

Introduction

"Quasar" or "Quasi stellar object" astronomical star-like objects having a large ultraviolet flux of radiation accompanied by generally broad emission lines and absorption lines in some cases found at large redshift. It is now known that quasars are distant but extremely luminous objects, so any light that reaches the Earth is redshifted due to the metric expansion of space Veron-Cetty and Veron 2010

Introduction(Cont...)

The taxonomy of quasars includes various subtypes representing subsets of featured application ion having distinct properties, the types of quasars are Radio-loud quasars, Radio-quiet quasars, Broad absorption-line (BAL) quasars, Type 2 (or Type II) quasars, Red quasars, Optically violent variable (OVV) quasars, Weak emission line quasars.



Finding the Optimal Number of Clusters

Dunn Index and Elbow Plot

Dunn's Validity Index Dunn† 1974 attempts to identify any number of clusters, where c_i represents the i -cluster of such partition, Dunn's validation index, D , can be calculated with the following formula:

$$D = \min_{1 \leq i \leq k} \left(\min_{i+1 \leq j \leq k} \left(\frac{\text{dist}(c_i, c_j)}{\min_{1 \leq l \leq k} \text{diam}(c_l)} \right) \right) \quad (1)$$

Elbow Plot It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used. It is the sum of the squared distances of samples to their closest cluster center.

▶ Go to graphs

▶ Go to Terms

Partitioning Methods and Classification

K means

K means clustering is a technique that aims to find more homogeneous subgroups within the data. The idea is to divide the Quasar data into k distinct groups so that observations within a group are similar, whilst observations between groups are as different as possible. Zahid Ansari 2011 Hartigan and Wong 1979

$$J(X, V) = \sum_{j=1}^k J_i(x_i, v_j) = \sum_{j=1}^k \left(\sum_{i=1}^m u_{ij} d^2(x_i, v_j) \right) \quad (2)$$

▶ [Go to Terms](#)

Partitioning Methods and Classification(Cont..)

Linear Discriminant Analysis

The **Linear Discriminant Analysis** technique is developed to transform the features into a lower dimensional space, which maximizes the ratio of the between-class variance to the within-class variance, thereby guaranteeing maximum class separability. The Linear Discriminant Function is $(\mu_1 - \mu_2)' \Sigma^{-1}$

▶ [Go to Terms](#)

Partitioning Methods and Classification(Cont..)

Xgboost

Xgboost: Tree boosting is a highly effective and widely used machine learning method. Here we describe a scalable end-to-end tree-boosting system called XGBoost. Chen and Guestrin 2016 Ghosh 2022. XGBoost works as Newton-Raphson in function space, a second-order Taylor Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = \arg \min_{\theta} \sum_{i=1}^N L(y_i, \theta) \quad (3)$$

the output will be

$$f_{(m)}(x) = f_{(M)}(x) = \sum_{m=0}^M \hat{f}_m(x)$$

▶ [Go to Terms](#)

Missing Value Treatment

The missing values in this data set have been imputed by the multiple imputation techniques known as Predictive mean matching(PMM). Basically, Predictive mean matching(PMM) calculates the predicted value of target variable Y according to the specified imputation model.

Results and Discussion

Result 1

Considering the clustering based on Lick indices(including the Redshift) we have observed 2 distinct clusters of size 129377 and 39851, and we came across an accuracy of about 99.16% and 99.85% when classified using the conventional multivariate technique Linear Discriminant Analysis and the ML Based classification Technique XGBoost.

Result 2

Similarly considering the heuristic literature-based partition(categorization) based on the redshift which is Low(0-2), Medium(2-4.1), and High(4.1-6.42) and applying the XGBoost classification algorithm an accuracy of about 95.92% is observed.





Conclusion



- From the affirmation results we can definitely conclude that the ML algorithms Perform at par with the conventional Multivariate Techniques, which are more common when handling astronomical databases.
- Combining the results of classification based on the lick indices(including redshift) and classification based on only redshift it is evident that even after providing extra information in the form of training data the accuracy of classification over the combined indices is greater. The train and test for K mean classification in 60% and 40% and for the redshift categorization, it is 80% and 20%.

Prospective Work

Since the Multivariate and Machine Learning techniques used in this work are not robust or absolute, so a possible comparison can be made between the different algorithms to discover the algorithm with the maximum efficacy. Additionally, a possible clustering cross-match can be done to infer the direction of probable variation in the photometric and physical properties of Quasars, corresponding to their redshift.

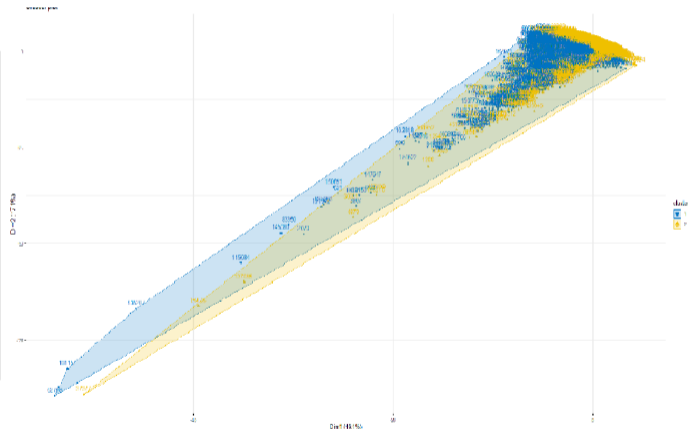
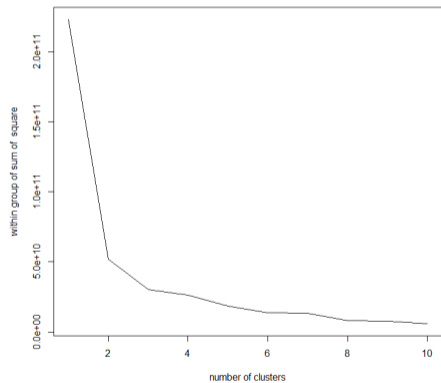
Thank You

-  Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.
-  Dunnt, J. C. (1974). "Well-Separated Clusters and Optimal Fuzzy Partitions". In: *Journal of Cybernetics* 4.1, pp. 95–104. DOI: 10.1080/01969727408546059. eprint: <https://doi.org/10.1080/01969727408546059>. URL: <https://doi.org/10.1080/01969727408546059>.
-  Ghosh, Prithwish (2022). "Breast Cancer Wisconsin (Diagnostic) Prediction". In: *International Journal of Science and research* 11.5, pp. 178–185. DOI: 10.21275/SR22501213650. URL: <http://dx.doi.org/10.21275/SR22501213650>.
-  Hartigan, J. A. and M. A. Wong (1979). "A k-means clustering algorithm". In: *Applied Statistics* 28, pp. 100–108.

-  Veron-Cetty, M. P. and P. Veron (July 2010). "A catalogue of quasars and active nuclei: 13th edition". In: *aap* 518, A10, A10. DOI: 10.1051/0004-6361/201014188.
-  Zahid Ansari M.F. Azeem, Waseem Ahmed (2011). "Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions". In: *World of Computer Science and Information Technology Journal (WCSIT)* 1, pp. 217-226. ISSN: 221-0741.

Appendix - A figure

[Return to presentation](#)



Appendix - Terms

Where $J_i(x_i, v_j) = \sum_{i=1}^m u_{ij} d^2(x_i, v_j)$, is the objective function within the cluster c_i , $u_{ij} = 1$, if $x_i \in c_j$ and 0 otherwise.

$d^2(x_i, v_j)$ is the distance between x_i and v_j $d^2(x_i, v_j) = \left\| \sum_{k=1}^n x_k^i - v_k^j \right\|^2$ where, n is the number of dimensions of each data point.

x_k^i is the value of k^{th} dimension of x_i , v_k^j is the value of k^{th} dimensions of v_j

The clusters are defined by a $m \times k$ binary membership matrix U, where the elements are u_{ij}

[Return to presentation](#)

Appendix - Terms

where $dist(c_i, c_j)$ is the distance between clusters c_i and c_j ,

$$dist(c_i, c_j) = \min_{x_i \in c_i, x_j \in c_j} d(x_i, x_j),$$

$d(x_i, x_j)$ is distance between data points $x_i \in c_i$ and $x_j \in c_j$,

$diam(c_l)$ is diameter of c_l where $diam(c_l) = \max_{x_{l_1}, x_{l_2}} d(x_{l_1}, x_{l_2})$

An optimal value of the k is one that maximizes Dunn's Index. index.

[◀ Return to presentation](#)

Dunn index for clustering

k	Number of Clusters = 2	Number of Clusters = 3	Number of clusters = 4
	18.03072757	-60.43461534	-7.093826221

[◀ Return to presentation](#)

Confusion Matrix for K means where $K = 2$, where accuracy = 99.16%

Actual — Predicted	Predicted Group 1	Predicted Group 2
Actual Group 1	38496	324
Actual Group 2	1085	129035

[Return to presentation](#)

Confusion Matrix for clustering by XGBoost, where accuracy = 99.85% with respect to the clusters after splitting the data in test(40%) and train(60%).

Predicted — Refrence	Reference Cluster1	Reference Cluster2
Predicted 1st Cluster	15784	48
Predicted 2nd Cluster	51	51692

[◀ Return to presentation](#)

Confusion Matrix for XGBoost, where accuracy = 95.92% with respect to the redshift

Predicted — Refrence Reference High Reference Medium Refrence Low

Predicted High	169	0	67
Predicted medium	0	26561	284
Predicted Low	26	1002	5678

[◀ Return to presentation](#)