computer sciences and mathematics forum

MDPI

Proceeding Paper

Spectral Classification of Quasar Subject to Redshift: A Statistical Study †

Prithwish Ghosh *,‡ and Shinjon Chakraborty ‡

Affiliation 1; shinjonchakraborty07gmail.com

* Correspondence: ghosh.prithwish1999@gmail.com; Tel.: +91-8961-312-165

† Presented at the 1st International Online Conference on Mathematics and Applications; Available online: https://iocma2023.sciforum.net/.

‡ The authors are having equal contribution to this work.

Abstract: Quasars are 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. The used data is extracted from Veron Cetti Catalogue of AGN and Quasar. The objective of this work is to partition the quasar based on their spectral properties using multivariate techniques and classify them with respect to the obtained clusters. Performing the K-means partitioning method two robust clusters were obtained with cluster sizes 39,581 and 129,377. The percentage of misclassification observed based on the obtained clusters considering a multivariate classification technique and machine learning classification algorithm i.e., "Linear Discriminant Analysis" and "XG-Boost" respectively. The Linear Discriminant Analysis and Xgboost evaluate a misclassification of around 0.84% and 0.15% Respectively. Additionally, a heuristic literature-based categorization subject to redshifts yields an accuracy of around 96%. It gives us cross-validating arguments about astronomical data, that machine learning algorithms might perform at par with the conventional multivariate techniques if not better.

Keywords: quasar; partitioning; K-means; XgBoost; discriminant analysis; machine learning

1. Introduction

"Quasar" astronomical objects like star with a large ultraviolet flux of radiation accompanied by generally broad emission lines and absorption lines in some cases found at large redshift. Nearly 10% of the quasars are radio-loud. The increment of redshifts for quasars can go up to z = 7 after which it decreases and there is a decrease to the higher redshifts. We known that quasars are extremely luminous objects and they are distant objects of our universe, so the lights which are able to reach the Earth, due to spaces metric expansion those are redshifted [1]. The supermassive black holes originate the power of quasars that are believed to exist at the core of galaxies. Near the cores of galaxies the Doppler shifts of stars which tells us that they are rotating around tremendous masses with very steep gravity gradients, suggesting black holes. The taxonomy of quasars includes various sub-types representing subsets of featured application ion having distinct properties, the types of quasars are **Radio-loud** , **Weak emission line quasars**, **Broad absorption-line (BAL)**, **Optically violent variable (OVV)**, **Radio-quiet**, **Type 2 (or Type II)**, **Red** quasars. Consisting of Parameters like Color indexes redshift, absolute magnitude, and magnitude. Assuming a deceleration parameter $q_0 = 0$, Hubble constant $H_0 = 71$ km/s/Mpc. The data set that we used consists of some parameters like The Declination of the object, The Right Ascension of the object, The (B-V) color of the object when known, The redshift of the object, The (U-B) color of the object, when known, etc.

check for updates

Citation: Ghosh, P.; Chakraborty, S. Spectral Classification of Quasar Subject to Redshift: A Statistical Study. *Comput. Sci. Math. Forum* **2023**, *1*, 0. https://doi.org/

Academic Editor: Firstname Lastname

Published: 28 April 2023

2. Materials and Methods

2.1. Missing Value Imputations

The censored values in this data set is been imputed by 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.

2.2. Choice of Optimal Clusters

2.2.1. Distortion Plot

The initial step for any unsupervised learning is to find out the optimal number of partitions in which the data may be partitioned. The Distortion plot Method is one of the most popular methods to determine this optimal value of k or the number of partitions to make. It is calculated as the average of the sum of squared distances from the partition centers of the created partitions. Here we used the Euclidean distance metric. It is the distance of samples to their closest cluster center with respect to their sum of the squared. Basically, where the curve or elbow-like part is given in the graph, it is the optimal cluster number from the graph is observed.

2.2.2. Dunn Index

Dunn's Index [2] tries to find those partitioned sets that are compact and well different. For a random number of partitions, where c_i represents the ith cluster, Dunn's index, D, is calculated with the formula given below:

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

where $dist(c_i, c_j)$ is the interval between clusters c_i, c_j,
$dist(c_i, c_j) = min_{x_i\in c_i, x_j\in c_j} dx_i, x_j$,
$d(x_i, x_j)$ is interval between data points $x_i \in c_i$ and $x_j \in c_j$,
$diam(c_l)$ is diameter of c_l where $diam(c_l) = max_{x_{l_1}, x_{l_2}} d(x_{l_1}, x_{l_2})$

The value which is optimal is responsible for Maximizing the Dunn's Index.

2.3. Clustering (Partitioning) Algorithms and Discriminant Analysis

Clustering is the technique of grouping individuals, having multiple characteristics, according to their similarities or dissimilarity. Basically, we are partitioning the data for getting information about how the Quasar data carries the information. Some of the pretty well-known algorithms which were used in the study are as follows-

2.3.1. K-Means

K means clustering which 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 [3,4].

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

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

x_k^i is the value of kth dimension of x_i, v_k^j is the value of kth dimensions of v_j

The partitions are defined by a m X k binary membership matrix U, where the elements are u_{ij}

$$u_{ij} = \begin{cases} 1 & \text{if } d^2(x_i, m_j) \leq d^2(x_i, m_{j^*}), j \neq j^*, \forall j^* = 1 \cdot k \\ 0 & otherwise \end{cases}$$

with fixed membership matrix U = $[u_{ij}]$, the choice of center v_j that minimizes J(X, V) is the mean for all the data's in the cluster j which can be calculated from the below equation:

$$v_j = \frac{1}{|c_j|} \sum_{i, x_i \in c_j}^{m} x - i \tag{3}$$

where $|c_j|$,which is the size of partitions c_j and also

$$|c_j| = \sum_{i=1}^{m} u_{ij}$$

2.3.2. The Linear Discriminant Analysis

To transform the features into a lower dimensional space, The Linear Discriminant Analysis technique is developed, where the ratio of the between-class variance to the within-class variance is maximized, which guarantees separability maximum class. The aim of the Linear Discriminant Analysis technique is to express the original data matrix onto a lower dimensional space. The Linear Discriminant Function is $(\mu_1 - \mu_2)' \Sigma^{-1}$.

2.3.3. XGBoost Algorithm

A highly used machine learning technique over various problems is Tree Boosting. Basically, the scalable end-to-end tree boosting system is called XGBoost [5]. In function space the XGBoost algorithm works as the Newton-Raphson method, A Taylor approximation of second order https://en.wikipedia.org/wiki/Taylor_series is used in to make the link to Newton Raphson Method(applied to the loss function). A generic unregulated XGBoost algorithm where we have to take input of a set (x_i, y_i) where i = 1 to N, a differentiable loss function L(y, F(x)),a learning rate α with a number of weak learners M.

Algorithm : Creating a model with value which is constant:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta) \tag{4}$$

for m = 1 to M. Then compute the gradient and Hessians:

$$\hat{g_m}(x_i) = \left[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)} \right]_{f(x) = \hat{f_{(m-1)}}(x)} \hat{h_m}(x_i) = \left[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)^2} \right]_{f(x) = \hat{f_{(m-1)}}(x)} \tag{5}$$

Now we have to fit a base learner with the help of training set $\left[x_i, -\frac{\hat{g_m}(x_i)}{\hat{h_m}(x_i)} \right]_{i=1}^{N}$ by calculating optimization problem we get that:

$$\hat{\phi_m} = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h_m}(x_i) \left[-\frac{\hat{g_m}(x_i)}{\hat{h_m}(x_i)} - \phi(x_i) \right]^2, \hat{f_m}(x) = \alpha \hat{\phi_m}(x) \tag{6}$$

the output will be

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

The features are **Speed**: It can automatically do parallel computation on Windows and Linux, with OpenMP. It is generally over 10 times faster than the classical game. **Input Type**: Several Types of input data can be taken here. **Dense Matrix**: R; s dense matrix, that is Matrix **Sparse Matrix**: R; s sparse matrix, that is Matrix or dgCMatrix **Data file**:

Local data's **Xgb.DMatrix**: its own class **Sparsity**: Both of the tree boosters and linear booster is accepted by this under sparse input. and the sparse input is also optimized. **Customization**: Customized objective function and evaluation functions are supported here [6].

3. Results and Discussion

Coming to the results after applying the aforementioned statistical techniques on the relevant dataset we observe that from the Elbow plot (Figure 1) and Dunn Index (Table 1) the optimal number of distinct clusters is two. After applying the k-means partitioning algorithm over the combined Lick indices, the robust clusters are thereby demonstrated in Figure 1. To check the extent of clustering i.e., the efficacy of the k-means the percentage of accuracy is calculated through the Linear Discriminant Analysis (a multivariate technique) and XG-boost (a Machine Learning algorithm). The accuracy percentage under Linear Discriminant Analysis is 96.16% and under XG-Boost is 99.85%, which are evident from Table 2 and Table 3 respectively. We have partitioned the data in two parts, defined as the train and the test data set, which are 60% and 40% respectively. Then we used the algorithm for the classification which is done on the test data set.

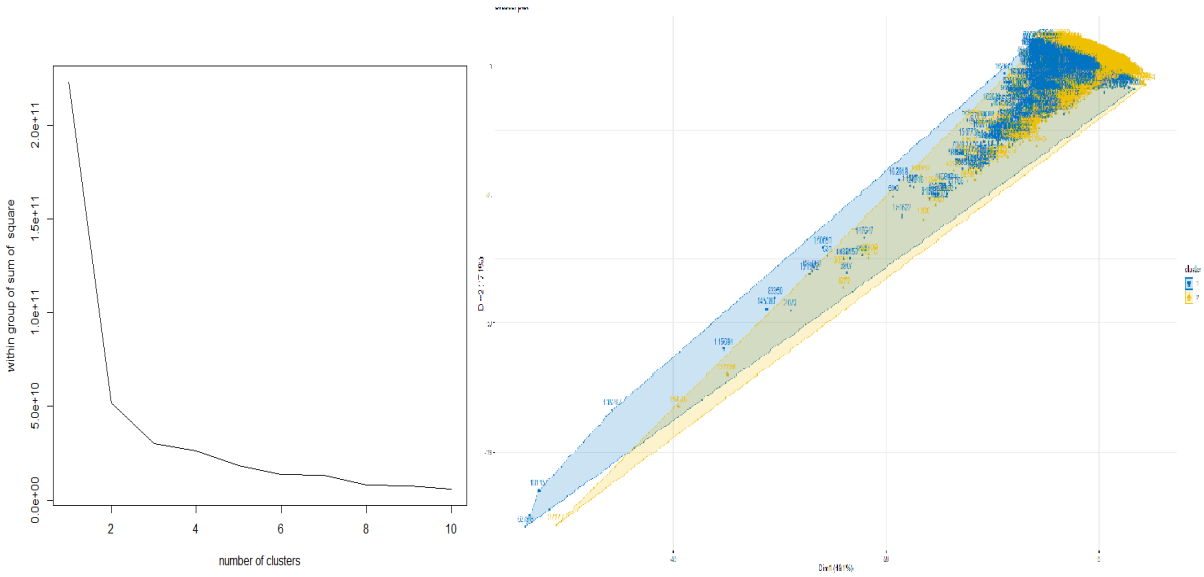

Figure 1. This is the figure for distortion curve and cluster plot.

Table 1. Dunn index for clustering.

k	Number of Partition = 2	Number of Partrition = 3	Number of Partition = 4
	18.03072757	−60.43461534	−7.093826221

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

Actual | Predicted	1st Predicted Group	2nd Predicted Group
1st Actual Group	38,496	324
2nd Actual Group	1085	129,035

Table 3. 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 Cluster 1	Reference Cluster 2
Predicted 1st Cluster	15,784	48
Predicted 2nd Cluster	51	51,692

We know that redshifts are one of the most complex and essential characteristics of an astronomical object since it is believed that properties of astronomical objects change with distance which is measured by the redshifts and also acts as a parameter in the spectral analysis. We have partitioned the quasars heuristically based on the redshifts into three distinct categories- Low redshifts (0–2), Medium redshifts (2–4.1) and High redshifts (4.1–6.44). We used the 80% and 20% partition for classification's train and test data. The classification table is shown in Table 4 and about 95.92% accuracy has been observed.

Table 4. 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	26,561	284
Predicted Low	26	1002	5678

4. Conclusions

Summarising the observations and results corresponding to a series of multivariate and machine learning techniques on a database of quasars from the Veron Cetti Catalogue (13th Edition), two distinct clusters are observed subject to a combination of lick indices (including redshift) designated as spectral properties of quasars with about 99% accuracy as computed by Linear Discriminant analysis and XG-boost algorithm.

Heuristically partitioning the quasars based on redshifts into 3 distinct groups viz. low, medium and high, it is observed that an accuracy of about 96% is encountered by applying the ML based classification algorithm XG-Boost which indicates the fact that the heuristic literature based clustering is quite valid in correspondance to the redshifts.

In context of astronomical studies we have mostly observed the utilisation of widely used multivariate techniques in majority of the studies. But we present an instance that the newly developed machine learning algorithms are quite on par with the wide used multivariate techniques, if not better.

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. So we may conclude in the direction that for Quasars clustering based on lick indices(including redshift) is more effective at least when working with the Veron Cetti catalog.

Author Contributions: Conceptualization, P.G. and S.C.; methodology, P.G. and S.C.; software, P.G.; validation, S.C.; formal analysis, S.C. and P.G.; investigation, P.G.; resources, S.C.; data curation, S.C. and P.G.; writing—original draft preparation, P.G.; writing—review and editing, S.C.; visualization, S.C. and P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement: The data set we are working on is extracted from Veron Cetti Catalogue [7] of AGN and Quasar (13th Edition). The dimension of this data set is 168,940 × 13. This database table contains the 13th edition of the Catalog of Quasars and Active Galactic Nuclei by Veron-Cetty and Veron. In this catalog contains 133,336 quasars, 1374 BL Lac objects, and 34,231 active galaxies (including 15,627 Seyfert 1 galaxies), for a grand total of 168,941 objects. It includes positions and redshifts, as well as photometry (U, B, and V) and 6-cm and 20-cm flux densities. https://heasarc.gsfc.nasa.gov/W3Browse/all/veroncat.html.

Acknowledgments: We are highly indebted to the renowned professors of Visva Bharati University and Calcutta University, for their constant support and motivation which resulted in the successful completion of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grupen, C.; Cowan, G.; Eidelman, S.; Stroh, T. *Astroparticle Physics*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 50.
2. Dunn, J.C. Well-Separated Clusters and Optimal Fuzzy Partition. *J. Cybern.* **1974**, *4*, 95–104.
3. Hartigan, J.A.; Wong, M.A. A k-means clustering algorithm. *Appl. Stat.* **1979**, *28*, 100–108.
4. Ansari, Z.; Azeem, M.F.; Ahmed, W.; Babu, A.V. Quantitative evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. *World Comput. Sci. Inf. Technol. J. (WCSIT)* **2011**, *1*, 217–226.
5. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
6. Ghosh Prithwish, L. Breast Cancer (Wisconsin) Diagnostic Prediction. *Int. J. Sci. Res.* **2021**, *11*, 178–185.
7. Veron-Cetty, M.P.; Veron, P. A catalogue of quasars and active nuclei. *Astron. Astrophys.* **2010**, *518*, A10