International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

Quasar Candidate Selection by Clustering using Fibonacci Series for Astronomical Surveys

Rohit K. Chandran[1], Ms. S. S. Pawar[2]

[1]PG Student, Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune affiliated to Savitribai Phule Pune University, Pune, India.

[2] Assistant Professor, Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Pune affiliated to Savitribai Phule Pune University, Pune, India.

Abstract

It is becoming apparent that the next generation astronomical analysis requires good domain knowledge to handle vast amount of astronomical data over network. The data availability over internet has increased tremendously and there exist a bridging gap between Computer Scientists and Astronomers. Most of the conventional methods are directly applied to the data without any simplification mechanism. The K-means is a clustering algorithm that has been used widely to classify dataset in astronomical databases. In this paper we propose an approach to group Quasar candidates with redshirts varying between 0.8 and 2.2 from SDSS catalogue. Most of these Quasars lie within a single region in a color versus red shift plot. For clustering, A similarity matrix is generated based on its photometric properties using their non-stellar colors ugriz photometry from SDSS Dataset. Clustering is performed on this matrix along with red shift. This matrix derived consists of score based on structured scientific information that leads to find interdependencies within the astronomical domain. The matrix eases the process of clustering by reducing number of attributes within each instance. The improvement in clustering performance can be understood when the inputs are of large scale.

Keywords: Data Mining, Astro-informatics, Clustering, Similarity evaluation, Fibonacci Series.

1.INTRODUCTION

WITH the advancement of technology and huge availability of dataset, efficient automated analysis has become an important issue for researchers working in astronomy. Massive datasets are released by various surveys, e.g. Sloan Digital Sky Survey(SDSS) which is one of the popular surveys of which five photometric bands (u', g', r', i', and z') cover the entire optical range from the UV atmospheric cut off at about 3000 °A to the red silicon sensitivity cutoff at about 10000 A° [2]. The SDSS consisted of imaging(photometric) data using a dedicated 2.5-m wide-angle optical telescope, and takes images using photometric system of five filters (named u, g, r, i and z). They have an important role in astronomy as it is the only information available about this distant object. In this system a process is adopted to group similar objects using an extended k-means algorithm. The extension is that each different attributes belonging to an instance of dataset is multiplied by a distinct successive Fibonacci value. Then the number of attributes in a large dataset is significantly reduced. Fibonacci series is used as a method to compute global score, such that ratio of two successive number converges to 1.6 or Φ, known as golden ratio. This matrix can be used to decide similarity among attribute while clustering. The matrix consists of an aggregate global score on scientific details in dataset like flux, magnitudes from photometric bands, red shift from spectrum, colour-colour plot, colour-magnitude plot, elliptical models. The Quasars are the most distant objects in the universe that can be observed. The emitted radiations from Quasars takes billions of years to reach the earth and therefore radiation gives information about the long ago state of quasar and about the early universe. The inverse-square law can be used to determine the luminosity of an object if its distance is known, simultaneously its distance can be determined if its luminosity is known. The photometric details can be used to generate flux using the Pogson magnitude function. They denote the brightness of the object with respect to the observer, lesser the magnitude brighter the object appears. The red shift can be used to determine whether the object is drifting away or moving towards us based on Doppler shift. The Balmer lines are useful in astronomy because in numerous stellar objects because abundance of hydrogen in the universe, and are commonly seen compared to lines from other elements. The study of these properties helps in basic understanding about the nature of the stellar objects. The predictions are made based on the cluster output combined with these scientific details.

2.RELATED WORK

In astronomy, petascale sky surveys will soon challenge our traditional research approaches and will radically transform how we train the next generation of astronomers, whose experiences with data are now increasingly more virtual (through online databases) than physical (through trips to mountaintop observatories)[1]. The problem of classifying and clustering quasars has already been addressed in the field of astronomy on photometric data as well as spectroscopic data, e.g., the idea given by Ball and Brunner. K- means algorithm is

the most commonly used clustering algorithm. This algorithm was developed to deal with instances of numerical attributes where distance between instances are a factor for clustering [3]. Most commonly used distance measures are Euclidean, Manhattan and Cosine. Automatic unsupervised classification of all SDSS/DR7 Galaxy Spectra has been developed to identify galaxies with similar spectra belonging to the same class using K-means [4]. Correlations among galaxy is used as an important clue for study of evolution of galaxies, here data is divided into two sub samples (blue and red samples) using a critical colour of (g-r) = 0.70 mag [5]. A simple step-by-step guide to qualitative interpretation of galaxy spectra is been developed for quick look analysis and for gaining physical insight in interpreting the outputs provided by automated tools. It compares the mean ages of the stellar populations with those inferred using a code STARLIGHT. A number of byproducts follow from the analysis. There is a tight correlation between the age of the stellar population and the metallicity of the gas, which is stronger than the correlations between galaxies mass vs stellar age, and galaxy mass vs. gas metallicity. The galaxy spectra are known to follow a one-dimensional sequence, and the luminosity-weighted mean stellar age as the affine parameter that describes the sequence is used for identification.[6]. A methodology for extracting knowledge from complex astronomical datasets using distinct unsupervised learning techniques is proposed by M. Brescia, S. Cavuoti, R. D'Abrusco, G. Longo and A. Mercurio in which an optimal clustering, based on a quantitative measure of the degree of correlation between the cluster membership is performed [7].In NASA's aeronautic and space missions, few of the most recent IT projects at Ames are focusing on innovative technologies for extracting information from astronomical data [8].

3.PROPOSED SYSTEM

The system is a layered architecture as shown in Figure.1.

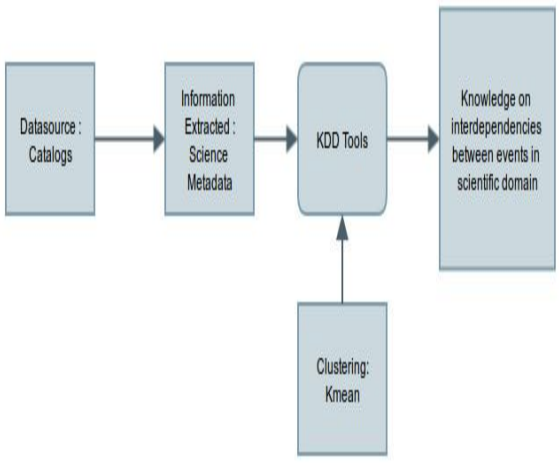

Figure 1 Quasar Selection System

Sky Survey data is collection of various photometric and Spectroscopic data and is the input to Informatics layer. The Informatics layer Implementation consists of input for analysis within KDD methods. This layer includes scientific meta data based on information extracted, also a score to create a clearly distinguishable space for easy selection. The scientific metadata included Pogson magnitudes and Balmer hydrogen lines for each instance. The data are further filtered using elimination from locus in color-color plot. The final output of the system is the prediction of inter dependencies within scientific domain. The Quasar selection metric consists of a global score defined on the basis of attributes like magnitudes, fluxes and red-shift. These features are considered as attributes of the object. The mathematical definition can be given by a set of instances $X = X_1 , X_2 , ..., X_n$ where X_i denotes 1 to n instances of astronomical dataset. These n instances can have m number of attributes. The global score reduces the search space from m to 1 where m is the number of attributes present in each instance.

$$R_n = \{(X_{n,1} , X_{n,2} , ...X_{n,m}\} \rightarrow \{(x_{n,1}\} \quad (1)$$

This $R_n = \{(x_{n,1}\}$ is used as similarity matrix. The following properties are used as the basis for generating selection, these factors remains closer for most of the objects.

a) Balmer Series: The hydrogen-containing region of spectrum can be obtained from red-shift using the formula:-

$$1 + z = \frac{lobserver}{lrest} \quad (2)$$

Here, $l_{observer}$ is wavelength with respect to observer and l_{rest} is rest wavelength. the observer wavelength for Hydrogen lines (Alpha. Beta, Gamma and Delta) can be calculated from redshift(z) and rest wavelength. Rest wavelengths of Hydrogen is as given in the following Table 1.

Table 1: Rest Wavelengths Of Hydrogen - BALMER SERIES

Name	Colour	Wavelengths (Angstroms)
Alpha (a)	Red	6562.8
Beta (b)	Blue-green	4861.3
Gamma (g)	Violet	4340.5
Delta (d)	Deep Violet	4101.7

b) Pogson Magnitudes: The Flux intensity can be calculated from magnitude for each u',g',r',i',z' using the formula:-

$$m = -2.5 \log \frac{f_x}{f_{x,0}} \quad (3)$$

Here, $f_{x,0}$ is the flux of the standard object taken (Vega Constellation), m is the apparent magnitude (u',g',r',i' and z') , flux f_x for u',g'.r',i' and z' can be calculated.

c) Colour-Colour Diagrams : The distribution of Stars and Quasars in Colour can be used as a factor for Quasar

selection , the region inhabited by stars should be removed first for Quasar selection. The colour of a ordinary star occupy a continuous one- dimensional region in (u'-g'),(g'-r'),(r'-i'),(i'-z') colour-colour space.(Newberg and Yanny 1997; Fan 1999; Finlator et al. 2000). Temperature is the primary parameter that decides position on stellar locus. Stars have a spectrum that is approximately black body in shape, quasars have spectra with blue continua and strong emission lines. The outlier in stellar locus can be identified as Quasars with certain exceptions.

4.IMPLEMENTATION DETAILS

Fibonacci series is used to find a global score for each instance and uses aggregate distance as a function to evaluate similarities . Fibonacci series can be defined as $F_n = F_{n-1} + F_{n-2}$. The first two terms are 0 and 1, and each successive number is the sum of the preceding two numbers. The property of Fibonacci series is that the ratio of two subsequent number F_n / F_{n-1} converges to 1.6, or φ known as the golden ratio [9]. The following methods are implemented:

Input Dataset: The datasets are fetched from SDSS using a Python API program which extracts features and labels as Comma Separate Values(CSV). The dataset includes u.g,r,i and z, also red shift. The objects are identified by their RA, DEC and objid. The CSV are loaded and necessary pre processing is performed.

Selection Approaches: The scientific value needed for implementation are calculated based on approaches described earlier. Flux value and Balmer series lines for hydrogen are calculated. These data remains as a basis for candidate selection. The colour-colour diagram used to identify the potential Quasars lying outside stellar locus of stars.

Colour - Colour Diagrams : The Colour- Colour diagram can be used as an excellent method for Quasar selection, Low-red shift quasars ($z < 2.0$) have blue u'-g' colours. They are lying differently from stars and white dwarfs in colour space and may misunderstood only by CELGs. The position of quasars at $z < 2.0$ changes little with red-shift, because of the dominant power-law continuum. For $z > 2$, the u'-g' colour becomes increasingly red. For $z > 3$, quasars are red in colours because of the absorption systems in their spectra. They are well separated from other kinds of stellar objects in colour space.

Colour Space: The colour space is another method that can be used for Quasar selection, according to Newberg and Yanny 1997; stars, galaxies and low red-shift quasars are distributed in the same"fundamental plane" in SDSS colour space. The stellar locus forms a ribbon-like structure in colour space. A set of stellar locus points are fitted to stellar distribution in the colour space of (u'-g', g'-r', r'-i'), for fitting by an ellipse the following set of axes are used :

- a1= 0.95 (u'-g') + 0.31 (g'-r') + 0.11 (r'-i')
- a2= 0.07 (u'-g') - 0.49 (g'-r') + 0.87 (r'-i')
- a3= -0.39 (u'-g') + 0.79 (g'-r') + 0.47 (r'-i')

a1 is along the average direction of the stellar locus for $T_{eff} > 4000K$, a2 is along the major axis of the fitted ellipse perpendicular to the stellar locus, and a3 is along the minor axis of the ellipse.

Score Matrice: An aggregate score is calculated for each instance applying a Fibonacci function.

1. $X = \{X_1 , X_2 , ..., X_n\}$;// Numerical instances with m attributes including color, red shift and their combinations $x_{i,1} , x_{i,2} , ..., x_{i,m}$.
2. $F = \{F_1, F_2, ..., F_m\}$; //Successive Fibonacci numbers F corresponding to each 1 to attributes.

The Fibonacci metric create a search space where input instances are easily separable. For n numeric instances and m number of attributes, the algorithm reduces search set for each $X = X_1, X_2,..... X_n$ from m to 1:

$$R\,n = \{(X_{n,1} , X_{n,2} , ... X_{n,m}\} \rightarrow \{(X_{n,1}\} \qquad (4)$$

A scaling factor z is used for normalizing attribute values. This is done to scale the values in a constant range so that Fibonacci number chosen for that attribute does not change the ratio much and make much difference.

$$score(xi) = \sum_{1}^{m} \frac{xij}{z} \cdot fi \qquad (5)$$

The instance similarities between X_i , X_j is calculated with the condition $Score(X_i) <= Score(X_j)$ and

$$Similarity(xi , xj) = \frac{score(xi)}{score(xj)} \qquad (6)$$

Output of Similarity Score:

$R_n = (x_{n,1})$ // Numeric instances : Global score

The above methods are implemented to classify the Quasars based on their properties.

5.DATASET

Our experiments work on a subset of the SDSS (DR6database. This data has been collected via a 2.5-meter Telescope at the Apache Point Observatory (New Mexico)along with two special-purpose instruments: a 120-mega pixel camera and a pair of spectrographs. This data is sent through FedEx to the master SDSS archive site at Fermilab in Illinois after each observing trial) [2]. The photometric and spectroscopic data are used in this selection process.

A. Photometric Data

In Sloan Digital Sky Survey(SDSS), a 120-megapixel camera collects data through five different filters. These data cover five wavelength ranges and are called u,g,r,i, and z bands. This photometric pipeline extracts data for all celestial objects. For each object, in each band, a set of photometric features are available. One type of these features, called magnitudes, are logarithmic measures of brightness of an object. Mostly used models for data fitting are the Petrosian, the PSF and the Model approach [11]. The datasets are retrieved from the PhotoObjAll table of the Catalog Archiver System(CAS) using SQL queried through an API . The following set of photometric features can be extracted from the five bands:

1) Apparent magnitudes: u, g, r, i and z.
2) PSF magnitudes: psfMag_u, psfMag_g,psfMag_r, psfMag_i, psfMag_z
3) Petrosian magnitudes: petroMag_u, petroMag_g, petroMag_r, petroMag_i, petroMag_z
4) Model magnitudes: modelMag_u, modelMag_g, modelMag_r, modelMag_i, modelMag_z

The datasets are based on the photometric and spectroscopic details of celestial objects taken in order to evaluate the performance of algorithm. The four S1-S5 datasets are based on different magnitude derived from the photometric data is shown in Table 2. These are various samples with varying red shift and magnitude with a specified range.

Table 2: CLUSTERING DATASET

Dataset	Attribute Type	No. of Attribute	No. of instance
S1	Numerical	5	2,056
S2	Numerical	5	5,012
S3	Numerical	5	10,000
S4	Numerical	5	20,000
S5	Numerical	5	30,000

The experiments were executed in python using AstroML[12]- a python based machine learning package and Scipy library for statistical estimation based on numpy, scipy, scikit-learn, and matplotlib.

6.RESULT AND DISCUSSION

K-means clustering is used with Fibonacci score to group the objects based on similarity using colour, Flux, Hydrogen emission lines using Balmer series and were grouped into different clusters. The cluster output generated consists of the mean centroid of each cluster. The output of the K-means and Fibonacci is shown in the Table 3.

Table 3: Centroid Table

Datapoint (N)	Elements clustered using K-means.	Elements clustered using Score K-means.
2095		
Cluster 0	868	798
Cluster 1	892	1052
Cluster 2	335	245
4096		
Cluster 0	589	448
Cluster 1	1807	2234
Cluster 2	1699	1413
8192		
Cluster 0	3608	4496
Cluster 1	3505	2879
Cluster 2	1080	822
10,000		
Cluster 0	4726	5050
Cluster 1	1371	854
Cluster 0	3903	4136
20,000		
Cluster 0	8669	6945
Cluster 1	8877	11205
Cluster 2	2454	1850
30,000		
Cluster 0	13232	10352
Cluster 1	13829	17545
Cluster 2	2939	2103

The colour-colour diagram and colour-magnitude diagrams used to further support the result. The scatter plot, color-color, color- red shift diagrams were used to support the result. The datasets selected were apparent magnitudes at different red-shift, concentrated on red-shift(Z_α) less than 2 or more than 1.0 . The shape of featureless continua of Quasars is taken approximately based on by power law (Vanden Berk et al. 2001). For the low-redshifted objects, In ugri colour space (magnitude $i \leq 19.1$), in the SDSS filters the Quasar locus is well separated from the stellar locus for relatively low($Z\ \alpha \leq 2$) and relatively high($Z\alpha \geq 2$) redshifts(Richards et al. 2001). If the colour of object is inside stellar locus(error region) the object is deemed to be non-Quasar and is rejected. If the object is outside surface, considered to be as good Quasar candidate as it is an outlier to stellar locus. For the high-red shifted objects, In griz colour space (magnitude $i \leq 20.2$), objects are not selected as Quasar candidates in griz when following conditions are satisfied

$$1)\ g* - r* < 1.0$$

$$2)\ u* - g* \geq 0.8$$

$$3)\ i* \geq 19.1 \text{or } u* - g* < 2.5$$

The score calculated based on different attributes were matching with the most of the condition identified for Quasars with high redshift an low redshift for the samples taken. The selections are accurate for the objects that are not close to the boundary of locus. The reddening may be a factor of variations in the certain properties that influenced the objects at the boundaries to drift from expected value. The speed up of execution was significant as the scoring reduced the complexity of dataset by reducing number of instances. The clustering were more of similarity search rather than finding distance between each instances. The execution time and speed up of random samples is as shown in the Table 4. The execution time calculated on random samples indicates that as the size of data increases score based method is a more viable solution for clustering.

Table 4: Execution Time

Datapoints(N)	*Time(sec) k-Means*	Time(sec) Score k-Means	Time Difference
100	0.17	0.10	0.07
512	l.5	0.90	0.6
1024	2.5	1.3	1.2
2048	4	2.1	1.9
4096	6	2.9	3.1
8192	9	4.3	4.7

7.CONCLUSION AND FUTURE SCOPE

In this paper, various aspects for Quasar selection are identified and on the basis of that a comparison metric is generated using novel idea of Fibonacci series. A similarity function is implemented to calculate the matching ratio between instances. The advantage of this method is that multiple attributes of several instances are modeled into a metric such that the selection process becomes easy and time saving. The search space is significantly reduced from simplifying n attributes to single attribute. The applicability of the methodology was Verified using colour-colour diagram and colour-magnitude diagrams. There are still many potential extensions along this line of research:

1) The selection can be extended to white dwarfs , as they have strong hydrogen lines; their colours can be modeled by a pure hydrogen atmosphere.

2) The selection of COMPACT EMISSION-LINE GALAXIES(CELGs) as they have power-law continua and strong emission lines, and thus their colours are similar to those of quasars with low-redshirts.

References

[1] "Sloan digital sky survey." http://www.sdss.org, 2010.

[2] K.D. Borne, Astroinformatics: data-oriented astronomy research and education.," Earth Science Informatics, vol. 3, no. 1-2, pp. 5-17, 2010.

[3] O. M. San, V. nam Huynh, and Y. Nakamori, "A alternative extension of the k-means algorithm for clustering categorical data," Int. J. Appl.Math. Comput. Sci, vol. 14, pp. 241–247, 2004.

[4] J. S. Almeida, J. A. L. Aguerri, C. Munoz-Tunon, and A. de Vicente, "Automatic unsupervised classification of all sdss/dr7 galaxy spectra," 2010.

[5] Z. Li and C. Mao, "Correlations among galaxy properties from the sloan digital sky survey," The Astrophysical Journal Supplement Series, vol. 207, no. 1, p. 8, 2013.

[6] J. S. Almeida, R. Terlevich, E. Terlevich, R. C. Fernandes, and A. B. Morales-Luis, "Qualitative interpretation of galaxy spectra," The Astrophysical Journal, vol. 756, no. 2, p. 163, 2012.

[7] M. Brescia, S. Cavuoti, R. D'Abrusco, G. Longo, and A. Mercurio, "Photometric redshifts for quasars in multi-band surveys," The Astrophysical Journal, vol. 772, no. 2, p. 140, 2013.

[8] M. Shafto and D. Korsmeyer, "Contributions to it: A view from ames research center," IT Professional, vol. 14, no. 2, pp. 13–19, 2012.

[9] R. Rawat, R. Nayak, Y. Li, and S. Alsaleh, "Aggregate distance based clustering using fibonacci series-fibclus," in Proceedings of the 13th Asia-Pacific Web Conference on Web Technologies and Applications, APWeb'11, (Berlin, Heidelberg), pp. 29–40, Springer-Verlag, 2011.

[10] Z. Wang, W.-G. Che, Y. Xiao, and C.-C. Yang, "Research of the Elliott wave theory applications based on cbr," in Intelligent System Design and Engineering Applications (ISDEA), 2013 Third International Conference on, pp. 1137–1140, 2013.

[11] A. Thakar, A. Szalay, G. Fekete, and J. Gray, "The catalog archive server database management system," Computing in Science Engineering, vol. 10, no. 1, pp. 30–37, 2008.

[12] J. VanderPlas, A. Connolly, Z. Ivezic, and A. Gray, "Introduction to astroml: Machine learning for astrophysics," in Intelligent Data Understanding (CIDU), 2012 Conference on, pp. 47–54, 2012.