RASCNA: Radio Astronomy Signal Classification through Neighborhood Assemblies

Mildred Morales-Xicohtencatl, Leticia Flores-Pulido,
Carolina Rocío Sánchez-Pérez, and Juan José Córdova-Zamorano

Autonomous University of Tlaxcala, Faculty of Engineering and Technology,
Road Apizaquito s/n. C.P. 90300 Apizaco, Tlaxcala, Mexico
http://www.uatx.mx/

Abstract. Computation is applicable to any branch in order to improve performance in times of process and results improvement, this article is the demonstration of an automatic process applied to the area of astronomy. The classification of electromagnetic spectra by pattern recognition is based on an assembly composed of neighborhood-based methods of classification. The acquisition of the electromagnetic spectrum to classify, is obtained of the SDSS III (*Sloan Digital Sky Survey*), the process of classification consists of a preprocessing, to obtain a specific region of the spectrum followed by filtering in advance of relevant features by means of digital signal processing and the wavelet haar transform. *abstract* environment.

Keywords: Electromagnetic spectrum, Quasar, Digital Processing of Signals, Wavelet Haar Transform, SDSS III.

1 Introduction

The astronomy radio studies and evolution of the universe, the celestial bodies as well as its natural emission of sound this study uses signals or electromagnetic waves applied to process in advance to the universe meaning through radio telescope signals. The study of these waves shows us essential features for the knowledge about the type of astronomical items classification. The range of the length of waves omitted from a body goes from 100 to 1000 *nm*, this range or spectrum usually belongs different regions radio waves frequency, microwave, infrared, visible light, ultraviolet and X-rays (see Figure 1).

This paper explains the study of electromagnetic signals to optimize the astronomic process of classification since it absorbs a large quantity of time for the astronomers differentiating the type of signal, luminosity, color, radial velocity, redshift, among others features required by automation of this process which becomes fundamental for a desired classification of data.

Section 2 an approach is proposed for the analysis of the signals. The proposal is applied at spectrum emitted from quasars which are the most distant objects and luminous of the universe, they are characterized by the extremely bright nucleus of an active galaxy, they are both in elliptical galaxies and spiral galaxies [5], [6].

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 222–230, 2014.

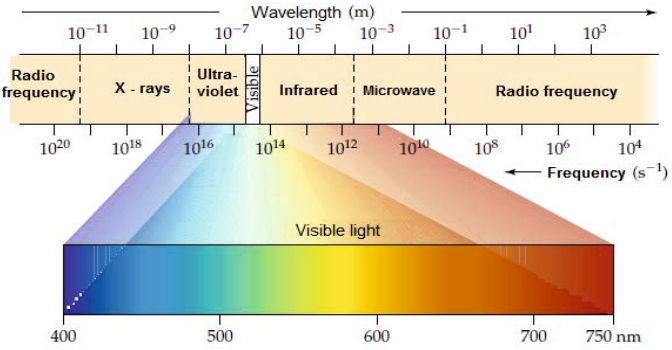

Fig. 1. Electromagnetic spectrum [1]

2 Acquisition of Astronomical Radio Data

The corpus of data used in this research consists of 500 signs which have been compiled from the SDSS III (*Sloan Digital Sky Survey*) [7], with a telescope belonging to the Apache Point observatory, in New México, this telescope has a primary lens of 2.5 *m*, a camera of 120 *megapixels*, a pair of spectrographs, each one made up of two cameras a red one and a blue one with a division of dichroic light the blue camera covers 3600 to 6350 *Angstroms*, the red camera covers 5650 to 10000 *Angstroms*. Each observation obtained by 1000 spectra from galaxies, quasars, and stars.

The spectrum used are acquired from the current version of data called DR9 (*Data Release 9*) from Studio BOSS (*Baryon Oscillation Spectroscopic Survey*)[2], since this compound of spectros belonging to galaxies and quasars in format Fits[1] (*Flexible Image Transport System*). The specifying of BOSS can be observed in Table 1 and Table 2.

Table 1. Data corresponding to DR9

SDSS III type of data in DR9	Amount of data
Objects of catalog	932, 891, 133
Spectra of Galaxies	1, 457, 002
Spectra of Quasares	668, 054

3 Processing of the Signal from a Quasar

The next phase of the processes will be described for the extraction of features in a signal emitted from a Quasar, apply to the signal different filtering and processes that can be observed in Figure 2 that shows the diagram of the stages that were applied to the signal for the feature extraction process.

[1] Fits, Main file format used in astronomy.

Table 2. Data corresponding to BOSS

BOSS
Observations made in Fall 2009 - Spring 2014
Wavelength: 360-1000 *nm*
Galaxies with the movement to the red of z = 0.7
Quasars with the movement to the red of 2.2 <z <3

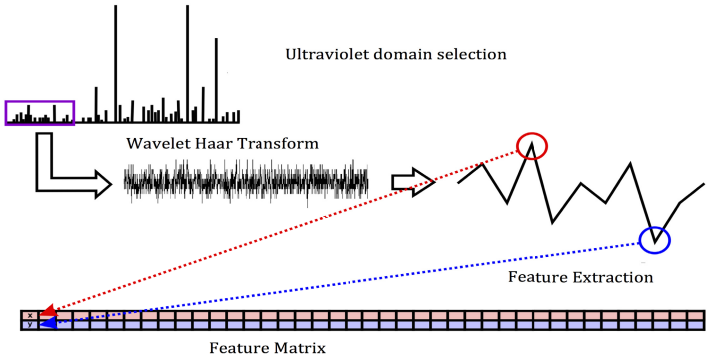

Fig. 2. Stages of the digital processing of signals

3.1 Reading and Filtration of the Ultraviolet Region

Once a time a spectrum is selected inside the data corpus, we proceeded at the fits archive reading as well as the identification of the type of signal included on it. The emission electromagnetic produced by the stellar bodies is an analog signal, which can be visualized as a matrix that contains information of the spectrum, the first filter applied to our signals if consists of the length of the spectrum corresponding to the ultraviolet region see Figure 3.

3.2 Application of the Haar Wavelet

Once ultraviolet region of the spectrum is filtered, thus, wavelet haar transform is applied to obtain a representation, decomposition and reconstruction of signals that show drastic changes in their time-frequency components (see equation (1)), as well as their representation in Figure 4.

$$s(t) = \begin{cases} 1 & 0 < t < 0.5, \\ -1 & 0.5 < t < 1, \\ 0 & \text{In another case} \end{cases} \tag{1}$$

Applying the Wavelet Haar Transform to the signal it is necessary to make a choice from a coefficient approach which contains the most representative values, see Figure 4 to view the levels of decomposition of the transformed Wavelet Haar, the decomposition level selected in the feature extraction process is of level five because, the energy preservation of the signal, es convenient for classification task.

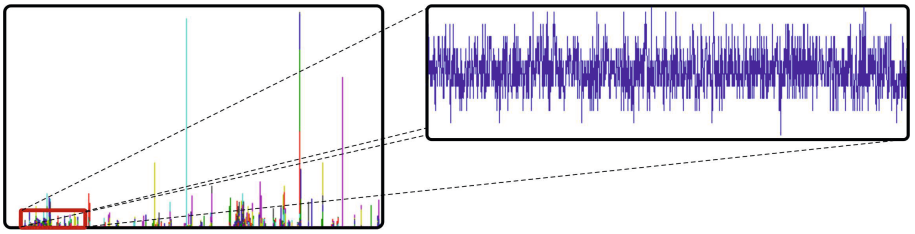

Fig. 3. Filtering of the ultraviolet region

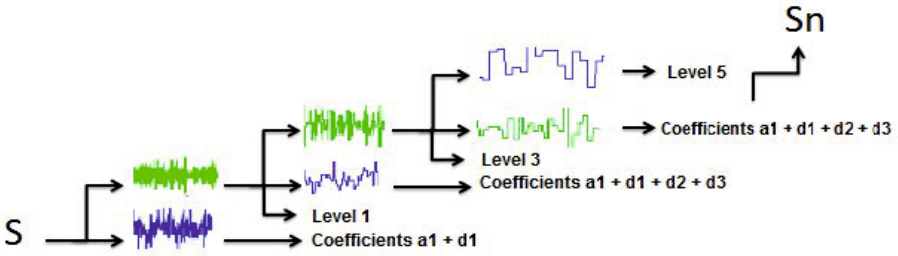

Fig. 4. Levels of decomposition

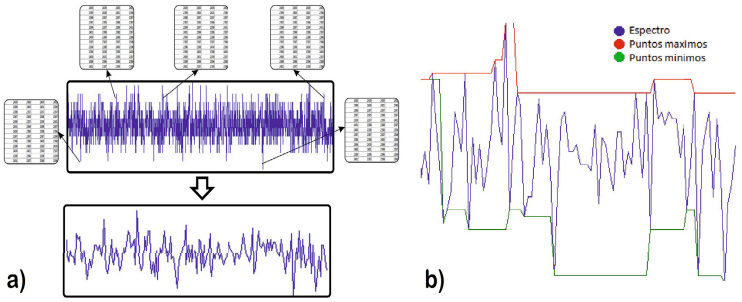

Fig. 5. a) shows the signal with a first filter corresponding to the values greater than 1000. b) we can to observe the extraction of features represented in minimums and maximums points.

3.3 Feature Extraction Task

The decomposition level is applied to the signal to filter and to obtain the most relevant values so they carry out two filters: the first one consists in obtaining the values higher than 1000 forming a new signal, then a secondfilter is applied that consists on removing 30 values obtained from the minimum and maximal points of each spectrum represented (see Figure 5 (b)).

Already computed the required filters a matrix of $2x30$ in computed too, where cleaned data are saved (see Figure 6).

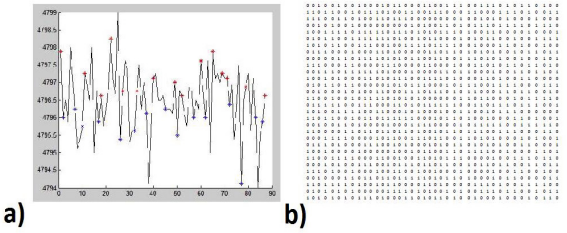

Fig. 6. a) shows the minimum and maximum values. b) Final matrix composed of m spectra.

3.4 Implementation of an Assembly Neighborhood-Based

The assembly is a set of 3 classifiers, the collection of two or more classifiers, in which their individual decisions are combined in order to obtain a more accurate result [3]. It is proposed the implementation of one assemble composed of methods of classification by neighborhood using an architectural hybrid, within which the classifiers are not related in such a way to classify a new case X a vote shall be obtained for each method of classification, (see Figure7).

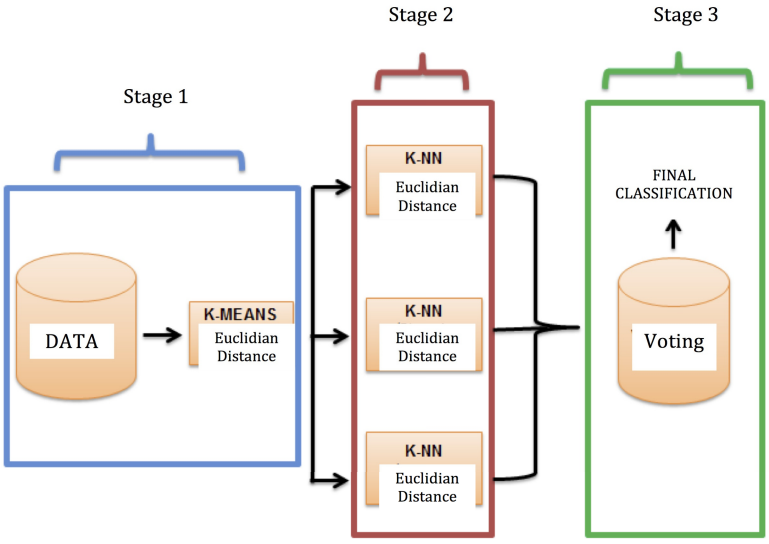

Fig. 7. RASCNA System: Signal Classification trough Radio Astronomy data Assembly

In step 1 of the Assembly (Figure 7) it implements an analysis of groups based on the algorithm k means which divides a data set in *Dn =(X1, X2,...,Xn)*. This analysis consists on the following steps:

- Step 1: assign the K desired groups, where each initial value of the center is equal to *Xi*, with the *Xn* objects belonging to *Dn*.
- Step 2: assigned data *X1* each of these points to the most proximate centroid of the centroids according to a measure of similarity.
- Step3: recalculate the centers of the k-groups.
- Step 4: repeat steps 2 and 3 until the reassignments are finished of the centroids.

For the operation of k-means there are taken 200 signals corresponding to 60% of the total existing data in the corpus of data, when taking these signals as an input for K-means the formation of three groups of data is obtained, Table 3 shows the distribution of data corresponding to each group.

Table 3. Distribution of samples by K-Means

Quasar (QSO)	Blazar (BL)	Radio Galaxy(RG)	Total Data
106	168	26	300

The result of the classification by grouping is presented graphically in Figure 8 which are classified in group 1 (BL), group 2 (QSO) and group 3 (RG).

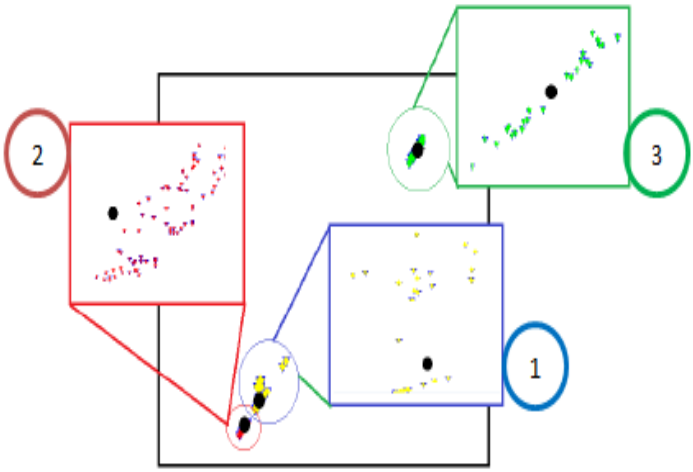

Fig. 8. Representation of groups in k-means

For the second stage of the assembly (Figure7), three sorters K-NN are implemented with different metric of similarity which are Euclidean, Manhattan and Murkowski [4]. K-NN is based on the assumption that the closest prototypes are the most similar in a way that this rule classifies X assigning the most representative tag between the closest K to the sample based on the following steps:

- Step 1: In a space D composed of samples already classified, for each sample X unclassified perform stages 2, 4.
- Step 2: Measure the distance to each element in the space.
- Step 3: The K elements are taken with the smallest distance.
- Step 4: The class for X element is assigned depending on the voting of the K elements assigned as their K-nearest neighbors.

Table 4 shows a comparative table of the three results obtained by the metric Euclidean, Manhattan and Minkowsky applied to a signal.

Table 4. Distances compared with similarity metrics Manhattan, Euclidean and Minkowski

Metrics	Manhattan	Euclidean	Minkowski
	Signal 1	Signal 2	Signal 3
X coordinate	4328.12	4328.12	4328.12
Y coordinate	4332.50	4332.50	4332.50
Distance	640.62	453.12	543.25
ID	117	117	117
Class	3	3	3

For the third and last stage a voting method has been implemented by majority, which is performed based on the individual votes by classifier.

In order to make this process more visual an interface has been implemented composed with the results of the most relevant processes applied by stages, to obtain a classification, in Figure 9, shows the interface of the system RASCNA system which is composed of three areas that corresponding to each stage of the assembly previously described in this section.

4 Analysis of Results

Table 5 shows the percentages of the spectral classification based on an Assembly composed of neighborhood classifiers, where it is observed how the results obtained by K-Medias and the method of votation and the method of voting of the ensemble complement each other, obtaining the best result in the Quasar class, followed by the Blazar class and with the worst percentage of classification in the group Radio Galaxy.

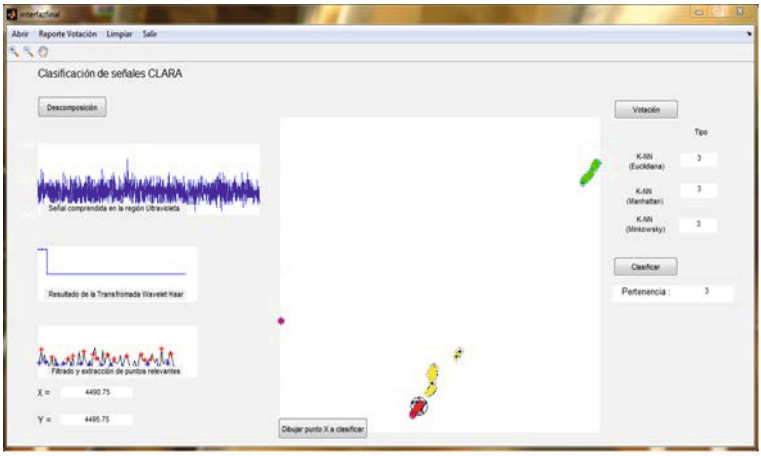

Fig. 9. User of interface of RASCNA system (Radio Astronomical Classification trough an Assemly)

Table 5. Classification rate obtained by K-means and the vote of the assembly

Method	Class Quásar	Class Blazar	Class Radio Galaxy
K-Means	87.23%	93.54%	50.98%
Vote assembly	91.07%	83.33%	65.00%
Percentage of classification	89.15%	88.43%	57.99%

5 Conclusion

The fulfillment of the procedures made and described throughout this article will allow us to extract the required features of the signal to identify the emission object type, allowing us to built a system that will the astronomical scientist the analysis of dependent procedures in a great part of the human factor in such a way that they obtain more optimal and punctual results.

For a classifier recognition it is essential to contain a set of features that mark decision limits, and to obtain a preprocessing of the signal like the most relevant values of each signal, taking these as coordinates for its position on a Cartesian plane. With these classification results of 89.15 % for the class QSO (Quasar), 88.43 % for BL (Blazar) and 57.99 % for RG (Radio Galaxies) that were obtained, it can be said that are acceptable for the two first classes however the class RG sheds a very low recognition, this is given by the margin of difference that exists in the landslide to the Red of the RG and BL class having values in X, Y very similar to these classes.

References

1. Serway, A., John, W., Jewett, Jr., Fisica, II.: Thomson, 3rd edn (2004)
2. Baryon Oscillation Spectroscopic Survey, http://www.sdss3.org/surveys/boss.php
3. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Unpublished Manuscript (1998)
4. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn., Canada (2000)
5. Gonzáles, G., Richards, J., Jay, W.: El planeta privilegiado. Palabra (2006)
6. Ruiz, J.: El universo en el III milenio. Sirius (2011)
7. Sloan Digital Sky Survey III, http://www.sdss3.org