Machine Learning approaches for detection and classification of astrochemical spectral lines

Alejandro Barrientos[ab], Mauricio Solar[b]
[a]Atacama Large Millimeter/submillimeter Array, Av. Alonso de Córdova 3107, Santiago, Chile
[b] Universidad Técnica Federico Santa María, Vicuna Mackenna 3939, San Joaquín, Santiago, Chile

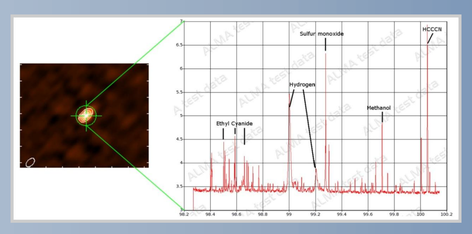
Abstract: Astronomical spectroscopy is a field that has been growing for a number of years, analyzing the features of molecular spectral lines from astronomical data cubes provides insight to the composition and dynamics of our universe. With the arrival of state-of-the-art high spectral resolution radiotelescopes like ALMA, the size of the data cubes will be constantly growing in time. This is why we believe that some automatic analysis methods will be helpful assisting the astrochemists work. We will generate a method to analyze astronomical data cubes, detect their regions of interest, by using a non supervised clustering algorithm, and then, generate a spectrum for each region of interest, and classify the molecular species found in the spectra, by using a supervised training algorithm. The training for the learner is done using synthetic spectra, and the validation is done using radio astronomical data cubes from ALMA observational data. A summary of related works is presented, and also a list of the astronomical complexities surrounding the nature of a molecular spectrum. Initial experiments contemplated a naive physical model that was considered to start the problem and two popular Machine Learning methods were tested for the task of classifying molecular spectra, Support Vector Machines and Neural Networks; results for SVM resulted in accuracy of over 90 percent with the basic model, later, a more complex model provided a slightly lower accuracy due to the lack of proper validation data. The Neural Network approach, provided similar results to the initial SVM approach. A parallelization test was also performed, obtaining a speedup of 2x in the process of real world data files.

Introduction

The study of the molecular composition of objects in space has been a matter of extensive research since the 1930's (Swings, Rosenfeld, 1937), through the use of Spectroscopy techniques. Around 150 molecules have been identified in interstellar or circumstellar clouds (Woon, 2015) and the number grows every year.

With the creation of new observatories like the Atacama Large Millimeter / Submillimeter Array (ALMA), a new window is being opened to space. Instruments with the sensitivity and resolution of ALMA will allow astronomers and astrochemists to further explore spectral line rich regions of space. ALMA observes at wavelengths in the range 3 mm to 400 m (84 to 720 GHz). With the high sensitivity of the observations, information about the chemical composition of the observing targets will be in the spectral lines captured by the telescope, some of them, for the first time. The analysis process to detect these molecules is quite complex and requires effort from astrochemists and laboratory spectroscopists (Cernicharo, 2012). Finding ways to classify some of the known lines is interesting, specially because in larger data cubes, thousands of lines can be present in a single detection, as ALMA can deliver up to 7680 frequency channels per data cube.
Our intention is to contribute with our efforts to the interdisciplinary field of Astro Informatics, by taking a computer science approach, namely the use of Datamining and Machine Learning algorithms and use them to learn to identify interesting emission regions in data cubes, and classify known spectra. Even though the specific process of identify and classify an astronomical spectrum is quite complex, we would like to start with a naive representation, and add complexity in future iterations.
The objective for this work is to train two well known machine learning models using synthetic data, and use such trained models with some of the ALMA Observational Data cubes to classify all of the existing trained molecules in the cube. We also want to include a parallelization variant, to make the process of data cubes faster.

Related Work



- GILDAS Weeds (Maret, 2015).
- CASSIS (Vastel, 2015).
- ADMIT (Teuben, 2015).
- MyXCLASS (Moller, et. al., 2013).
- MADEX (Cernicharo, et,al., 2012).
- Source Extractor (Bertin, et. al., 1996).
- ClumpFind (Williams, et. al., 1994).

- Molecular Line Association Analysis (Miranda, 2015).
- Detection and classification of spectroscopic lines (Pichara, 2013).
- A Machine Learning Application for Classification of Chemical Spectra (Madden, 2009).
- Combining Genetic Algorithms, Neural Networks and Wavelet Transforms for Analysis of Raman Spectra (Hennesey, 2004).
- Indexing data cubes for content-based searches in radio astronomy (Araya, 2016).
- Knowledge Discovery in Mega-Spectra Archive (Skoda, 2015).

Astronomical Complexities

There are several astrophysical complexities that surround the problem of line analysis (Tennyson, 2005)

Source Complexities: Inherent to the astronomical source that is being observed and its nearby regions. Examples: Temperature, Optical Depth, Abundance, Critical Density, Source Kinematics, Pressure.

Molecular Complexities: Inherent to the molecules in the observed source. Examples : Abundance, Internal Structure, Electronic Transitions, Molecular Rotation, Vibrational Rotation, Frequency.
Other factors: Line Shape, Literature is not always precise.

Objectives

Having established that there are two interesting problems:
1. Detection of regions of interesting emission in Astronomical Data Cubes.
2. Manual classification of astrochemical spectra, which can be a complicated task due to several factors that affect the characteristics of a spectral line on their way towards us.

What do we want:
1. To implement an algorithm that allows us to find out regions with significant energy emission from the sources, using clusterization techniques. Try several, choose the best one.
2. For each region, classify known spectral lines by using one supervised classification Machine Learning model. Try several and choose the best one (criteria: Accuracy vs. ease of implementation)
3. If time allows, try parallelization to make it faster (Optional)

Detection of Regions of Interest (ROI)

Method Selection : DBSCAN, Current parameters are: Epsilon : 10% of FITS size. MinPts = 7. Algorithm: Compact data cube into 2D, crop data below 3-Sigma, run clusterization. For each cluster, generate spectrum.
Pros: Irregular shapes can be detected no prior knowledge of number of clusters is required, unlike globular methods like k-means.
Cons: Very sensitive to parameter selection, need to find an automatic way to select them. Current ROI algorithm only works for high emission. Low emission is lost in noise when collapsing cube.

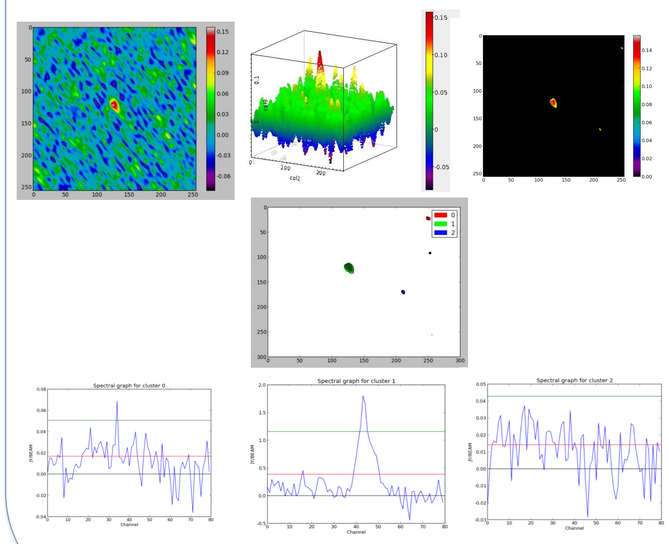


Figure 1, Processing of FITS file to detect regions of interest via DBSCAN

Artificial Neural Networks (ANN)

Implementation (Barrientos, Flores, Gonzalez, 2015): The original approach considered 8 transitions, with 100 examples per transition, each training example included intensity information for 1000 channels where the highest intensity was at the center frequency.
These models were created using the library Pybrain (Schaul, 2010). A feed-forward network with sigmoidal activation functions was used. For the output layer, a softmax function was used. For the training phase, backpropagation was used. As a measure of training error, we used the default pybrain error - Mean Squared Error - of the output vector in relation to the vector of the corresponding class.

Model Name	Description
M1_100	Model with 1 hidden layer with 100 neurons
M2_10	Model with 1 hidden layer with 10 neurons
M3	Model without hidden layer

Table 1: ANN Models

M1 obtained a validation error close to 0.01% while M2 obtained an error of 0.15%. M3 obtained the lowest error with under 0.002%. The data was divided 80% for training and 20% for testing. For validation, 10% of the training data was used.

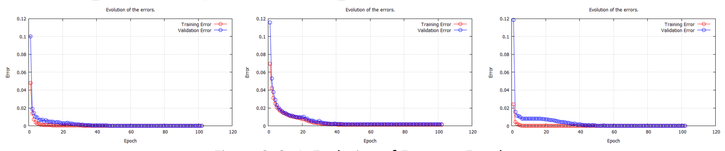


Figure 2, 3, 4: Evolution of Errors vs Epoch.

Model	Time(min.)	Epochs	Training Error	At Epoch	Test Error	Missclasification Error
M1_100	328	102	0.005297%	98	0.019398%	0%
M2_10	36	102	0.122776%	102	0.151988%	0%
M3	29	102	0.000076%	91	0.003531%	0%

Table 2: ANN Testing Results

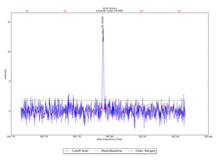


Figure 5, CO (3-2) Training Example, modified ADMIT code

Support Vector Machines (SVM)

Implementation (Barrientos, Ferreira, 2015): The SVM approach was implemented using scikit-learn (Pedregosa, 2011), libSVM-gpu (Chang et.al. 2011). The workflow follows these steps.

The first test consisted on detecting single transitions as standalone classes, while this approach is very naive, it was a good starting point. The classification accuracy is described in table

Threshold	2.5	3	3.5	4	4.5	5
Accuracy	0.91479	0.93032	0.94032	0.94285	0.94739	0.94505

Table 3, SVM Results, First Approach.

The second approach consisted on detecting “all” transitions in ALMA range (84-940 GHz) for a given species, to consider classification. In other words, the more transitions existed, the higher the probability of classification into that particular species. For classification we tested both One-vs-One and One-vs-All approaches, both provided similar results.

Threshold	2	2.5	3	3.5	4	4.5	5
Accuracy	0.8626	0.8706	0.8784	0.8835	0.8741	0.8887	0.8938

Table 4, SVM Results, Second Approach.

Extra Mile, Parallelization

To increase the performance, a few tests were made where the ROI Detection algorithm was modified to a parallel version where you could process “n” number of fits files at the same time using “n” processors. The implementation was made using mpi4py library.

Files	1	2	3	4	5	6	7	8
Sequential	6.445	51.828	58.162	64.944	77.369	85.453	170.117	217.807
Parallel	6.814	44.214	48.7	45.164	51.628	49.964	89.644	97.763
Speedup	0.946	1.172	1.194	1.438	1.499	1.710	1.898	**2.228**

Table 5, Parallelization test on ROI algorithm

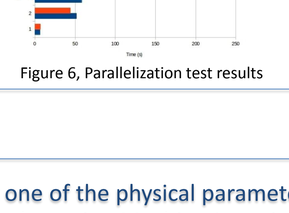


Figure 6, Parallelization test results

Conclusions

The problem of classifying astronomical spectra contains many extra variables which make it a more complex problem than to classify laboratory spectra, In this first approach, just one of the physical parameters has been taken into consideration (frequency) to classify the different spectra, we are aware that the current implementation is not physically accurate to the “real world” model, however we believe that using Machine Learning on these early stages will help to lay a foundation for more complex projects of this kind, our second approach using SVM showed lower accuracy, although this is expected because the higher complexity of the new model required a new training set. We expect to develop new approaches that allow us to keep or improve the accuracy of the models, while incorporating more astrochemical concepts into them. The scarcity of real world training examples makes the synthetic data useful to attempt further refinements.
In the parallelization testing, it was possible to develop a parallel python program using mpi4py that is capable to process 8 data cubes of 660 MB in size of dimension 300x300x1920 in 1.917 seconds, 3.87 times faster that its sequential counterpart. Applying these principles to an existing program that uses 8 FITS files of different sizes, it was possible to achieve a 2.22x speedup.

REFERENCES
- Araya, M., Candia, G., Gregorio, R., Mendoza, M., Solar, M., 2016. Indexing data cubes for content based searches in radio astronomy. Astronomy and Computing 14, 23 – 34. URL http://www.sciencedirect.com/science/article/pii/S2213133716000032
- Araya, M., Mardones, D., Hochfarber, T., 2015. Exorcising the ghost in the machine: Synthetic spectral data cubes for assessing big data algorithms. Astronomical Data Analysis Software and Systems XXIV (ADASS XXIV) 495, 57–60.
- Bertin, E., Arnouts, S., Jun. 1996. SExtractor: Software for source extraction. Astronomy and Astrophysics Supplement 117, 393–404.
- Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.
- Born, M., Oppenheimer, R., 1927. Zur quantentheorie der molekeln. j-ANN-PHYS-1900 389 (20), 457–484.
- Cernicharo, J., 2012. Laboratory astrophysics and astrochemistry in the herschel/alma era. http://tinyurl.com/ogf562e.
- Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
- Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning 20, 273–297.
- Hennessy, K., Madden, M., Ryder, A., 2004. Combining genetic algorithms, neural networks and wavelet transforms for analysis of raman spectra. URL https://www.analyzeiq.com/publications/aics2004a-paper.pdf
- Holwerda, B., 2016. Source extractor for dummies. http://tinyurl.com/jkghons.
- Madden, M. G., Howley, T., 2009. Applications and Innovations in Intelligent Systems XVI: Proceedings of AI-2008, the Twenty-eighth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer London, London, Ch. A Machine Learning Application for Classification of Chemical Spectra, pp. 77–90. URL http://dx.doi.org/10.1007/978-1-84882-215-3_6
- Maret, Sebastien, e., 2015. Weeds. http://tinyurl.com/ot6n9ll.
- Miranda, N., Cabrera, G., 2015. Association rules for spectral lines. URL http://repositorio.uchile.cl/handle/2250/133022
- Moller, T., Schilke, P., 2015. Manual for xclass-interface. https://www.astro.uni-koeln.de/wd-schilke/myXCLASS/XCLASS-Interface__No-NR__Linux__version_1.1.6.zip.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.
- Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408.
- Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Ruckstieß, T., Schmidhuber, J.,2010. PyBrain. Journal of Machine Learning Research 11, 743–746.
- Schilke, P., Moller, T., Comito, C., Sanchez-Monge, A.,Schmiedeke, A., Zernickel, A., Dec. 2015. Taming the Dragon: Automatic Line Fitting of ALMA data. In: Iono, D., Tatematsu, K., Wootten, A., Testi,L. (Eds.), Revolution in Astronomy with ALMA: The Third Year. Vol. 499 of Astronomical Society of the Pacific Conference Series. p. 195.
- Swings, P., Rosenfeld, L., 1937. Considerations regarding interstellar molecules. Astrophysical Journal 86, 483–486.
- Tennyson, J., 2005. Astronomical Spectroscopy: An Introduction to the Atomic and Molecular Physics of Astronomical Spectra. Imperial College Press.
- Teuben, P., 2015. Admit 0.5.2 documentation. http://tinyurl.com/osbu7ss.
- Skoda,P., Bromova, P., Lopatovsky, L., Palicka, A., Vavzmy, J., Sep. 2015. Knowledge Discovery in Mega-Spectra Archives. In: Taylor, A. R., Rosolowsky, E. (Eds.), Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV). Vol. 495 of Astronomical Society of the Pacific Conference Series. p. 87.
- Vastel, C., Bottinelli, S., Caux, E., Glorian, J.-M., Boiziot, M., Dec. 2015. CASSIS: a tool to visualize and analyse instrumental and synthetic spectra. In: Martins, F., Boissier, S., Buat, V., Cambresy, L.,Petit, P. (Eds.), SF2A-2015: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics. Eds.: F. Martins, S. Boissier, V. Buat, L. Cambresy, P. Petit, pp.313-316. pp. 313–316.
- Williams, J. P., de Geus, E. J., Blitz, L., Jun. 1994. Determining structure in molecular clouds. Astrophysical Journal 428, 693–712.
- Woon, D., 2015. Lists of interstellar and circumstellar molecules. http://tinyurl.com/m69ll6qg.

In search of our Cosmic Origins