Development of an Astro-Informatics Platform for Management and Intelligent Analysis of Large-scale Data

"Molecular Lines Association Rules"

Guillermo Cabrera.

University of Chile, November 26, 2013

Abstract

Historically, spectral surveys have been created by analyzing one spectra at the time. ALMA data-cubes offers new opportunities for automated spectral lines detection, as we will have spectral data for a considerable amount of pixels within each observation. This poses new challenges for data-mining tools. This project aims to study and apply association rule learning algorithms to ALMA data-cubes in order to classify spectral lines and find new relations between unknown lines and molecular transitions. We will create a framework for automatically process ALMA data-cubes in a high performance computing environment.

Keywords: Data Mining, Association Analysis, Association Rules, Unsupervised Learning.

Contents

1 Introduction **2**

1.1 Association Rules 2

1.2 Previous Work 2

2 Methodology **3**

3 Work Plan **3**

3.1 Activities 3

References **4**

1 Introduction

The Atacama Large Millimeter/sub-millimeter Array (ALMA) is an astronomical interferometer consisting of 66 antennas observing at millimeter and sub-millimeter wavelengths. ALMA produces an important amount of data in the form of the so called *data-cubes*. Each of these data-cubes is a 3D image where the first two dimensions are angular coordinates, and the third dimension is frequency. In other words, for each pixel of these images we have a whole spectrum.

This huge amount of data poses new challenges and opportunities to discover new physics. In particular, an important problem that ALMA data will allow to address is that of creating molecular images, associate new molecular lines to existing set of lines, and even identify new transition lines.

Our work will focus on creating association rules for molecular lines in ALMA data-cubes. These association rules will allow us to create images for different transitions, and associate unclassified molecular lines to existing or new transitions.

1.1 Association Rules

Association rule learning (Agrawal et al., 1993, ARL) is a popular data mining tool for discovering relations between variables in large data-sets. Consider a set of p variables $\mathcal{X} = \{X_j\}_{j=1}^{p}$. Usually, these variables are taken to be binary: $X_j \in \{0, 1\}$. We define a set of N *transactions* $\mathcal{D} = \{\boldsymbol{t}_i\}_{i=1}^{N}$, where $\boldsymbol{t}_i = \{x_{i,j}\}_{j=1}^{p}$, and

$$x_{i,j} = \begin{cases} 1 & \text{if item } j \text{ is part of transaction } i \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The main goal of association rule analysis is to obtain joint values of variables $(X_1, X_2, \cdots, X_p)$ that appear more frequently on our data-set.

For the case of molecular bands association rules, we will use lines detected at different pixels in different images as our transactions. In other words,

$$x_{i,j} = \begin{cases} 1 & \text{if a line is detected at frequency } \nu_j \text{ in pixel } i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Using common association analysis algorithms, such as the *Apriori* (Agrawal et al., 1994) and *FP-growth* (Han et al., 2000) algorithms, we will find sets of lines that jointly appear in the data with high probability. These association rules will be further used to detect molecular transitions and associate unknown lines with existing molecules and even distinguish new ones.

1.2 Previous Work

Spectral line detection has usually been made on a one-spectrum basis (Schilke et al., 2001; Watanabe et al., 2012; Cernicharo et al., 2013). The procedure usually consists of sideband deconvolution, relative calibration or pointing offsets, and identifying lines from catalogs (see Schilke et al., 2011, and references therein).

Different catalogs has been created for the purpose of line detection, including CDMS (Müller et al., 2005), JPL (Pickett et al., 1998), Lovas/NIST (Lovas, 2004), ToyaMA (Kobayashi et al., 2013), and OSU (Medvedev & Fortman, 2013). These catalogs contain detailed description of known molecular transitions. SPLATALOGUE (Remijan & Markwick-Kemper, 2008; Remijan, 2010) is a state of the art catalogue that combines different catalogs into a single database, which is publicly available for quering.

Association rule learning (ARL) is a widely developed field in machine learning. Rule learning was introduced by Piatetsky-Shapiro (1991) and formerly developed as association rule learning by Agrawal et al. (1993). Most popular ARL have been reviewed and compared by Hipp et al. (2000) and Ceglar & Roddick (2006). Though it has been vastly used for problems such as market basket analysis (see, for example, Chen et al., 2005) to the best of our knowledge it has never been applied to molecular lines detection.

2 Methodology

Our main goal is to create an association rule learning tool for associating spectroscopic lines to molecular transitions using ALMA data-cubes. We will try our algorithms over simulated and real ALMA data, and allow this tool to work on an HPC environment.

We will create simulated ALMA data-cubes based on the lines catalog of SPLATALOGUE. For this purpose, we will create simulated images following ALMA resolution for different molecules, and will add this information at frequencies of lines associated to these molecules. As a second simulation step, we will associate fake lines to these molecules, which will take into account unknown lines that may be associated to existing molecules.

Different ARL algorithms will be studied. We will evaluate which of these algorithms may be applied to ALMA data-cubes and program them in order for later applying each of them over our set of mock data-cubes.

Different metrics for comparing results with different methods will be defined. These metrics will be statistically defined taking into account maximum likelihood and a posteriori probabilities.

Finally, all methods will be tested over real ALMA data-cubes. This data will be obtained from publicly available ALMA data.

3 Work Plan

3.1 Activities

Main activities to be done during this project will be:

1. **Problem Definition and Candidate Selection:** We will formally define the problem, create a thesis project, and search for a student to work on this topic.

   1.1 Problem Definition

1.2 Thesis Project Announcement

1.3 Candidate Selection

2. Understanding the Data: During this activity we will understand how data-cubes are represented, and how SPLATALOGUE is organized.

2.1 ALMA Data-Cubes: Understanding ALMA data-cubes.

2.2 SPLATALOGUE: Understanding SPLATALOGUE.

3. Tool for Mock Data-Cubes: During this activity we will develop the tool for creating simulated ALMA data-cubes.

3.1 Single Frequency Image: Creating a simulated 2D image for a particular frequency.

3.2 Extending to Multiple Frequencies: Duplicating previous image to multiple frecuencies for a particular molecule based on SPLATALOGUE.

4. ARL Tool: Development of an ARL tool for obtaining associated lines sets.

4.1 Study ARL Algorithms: Evaluate which ARL is appropriated for this particular problem..

4.2 Development of ARL Tool: Create an ARL tool for ALMA data.

4.3 Evaluation Metrics: Development of metrics for assessing our methods.

5. Testing: We will test our methods over our mock and real ALMA data-cubes.

5.1 Testing over Mock Images

5.2 Testing over Real Images

5.3 Writing Study Report: We will compare our results with different methods and write a detailed report about over which images each one works better.

Figure 1 shows a Gantt chart describing how these activities will be executed during the 16 month period that this project will last.

Main Activity	Specific Activity	Aug. 2013	Sept. 2013	Oct. 2013	Nov. 2013	Dec. 2013	Jan. 2014	Feb. 2014	Mar. 2014	May 2014	Jun. 2014	Jul. 2014	Aug. 2014	Sept. 2014	Oct. 2014	Nov. 2014	Dec. 2014
1. Problem Definition and Candidate Selection	1.1 Problem Definition							X	X								
	1.2 T. Project Announcement							X	X								
	1.3 Candidate Selection							X	X								
2. Understanging Data	2.1 ALMA Data-Cubes							X	X								
	2.2 SPLATALOGUE							X	X								
3. Tool for Mock Data-Cubes	3.1 Single Frequency Image							X	X								
	3.2 Ext. To Multiple Frequency							X	X								
4. ARL Tool	4.1 Study ARL Algorithms							X	X								
	4.2 Development of ARL Tool							X	X								
	4.3 Evaluation Metrics							X	X								
5. Testing	5.1 Testing over Mock Images							X	X								
	5.2 Testing over Real Images							X	X								
	5.3 Writing Study Report							X	X								

Figure 1: Gantt chart of project activities.

All work will be done by an engineering or master level student as part of his thesis. He will spend at least 30 hours per week for this project.

References

Agrawal, R., Imieliński, T., & Swami, A. 1993in , ACM, 207–216

Agrawal, R., Srikant, R., et al. 1994, in Proc. 20th Int. Conf. Very Large Data Bases, VLDB, Vol. 1215, 487–499

Ceglar, A., & Roddick, J. F. 2006, ACM Computing Surveys (CSUR), 38, 5

Cernicharo, J., et al. 2013, The Astrophysical Journal Letters, 771, L10

Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. 2005, Decision support systems, 40, 339

Han, J., Pei, J., & Yin, Y. 2000in , ACM, 1–12

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. 2000, ACM SIGKDD Explorations Newsletter, 2, 58

Kobayashi, K., Tsunekawa, S., Matsushima, F., & Ohishi, M. 2013, Toyama Microwave Atlas for spectroscopists and astronomers

Lovas, F. J. 2004, Journal of Physical and Chemical Reference Data, 33, 177

Medvedev, I., & Fortman, S. 2013, The Ohio State University Data contained in Splatalogue

Müller, H. S., Schlöder, F., Stutzki, J., & Winnewisser, G. 2005, Journal of molecular structure, 742, 215

Piatetsky-Shapiro, G. 1991, Knowledge discovery in databases, 229

Pickett, H., Poynter, R., Cohen, E., Delitsky, M., Pearson, J., & Müller, H. 1998, Journal of Quantitative Spectroscopy and Radiative Transfer, 60, 883

Remijan, A. J. 2010, in Bulletin of the American Astronomical Society, Vol. 42, 568

Remijan, A. J., & Markwick-Kemper, A. J. 2008

Schilke, P., Benford, D., Hunter, T., Lis, D., & Phillips, T. 2001, The Astrophysical Journal Supplement Series, 132, 281

Schilke, P., Rolffs, R., & Comito, C. 2011, Proceedings of the International Astronomical Union, 7, 440

Watanabe, Y., Sakai, N., Lindberg, J. E., Jørgensen, J. K., Bisschop, S. E., & Yamamoto, S. 2012, The Astrophysical Journal, 745, 126