*Desarrollo de una plataforma astroinformática para la administración y análisis inteligente de datos a gran escala*

# Detección, Caracterización y Clasificación de Líneas Espectrales Utilizando Descriptores Estadísticos

Mauricio Solar, Marcelo Mendoza, Jonathan Antognini
Andrés Riveros, Diego Mardones, Guillermo Cabrera,
Karim Pichara, Nelson Padilla, Ricardo Contreras,
Neil Nagar, Victor Parada.

Santiago, 6 de noviembre de 2013

**Resumen**

La detección y caracterización automática de líneas espectrales es un problema astronómico que no ha sido abordado hasta ahora. A través del uso de la espectroscopía es posible describir la composición química de objetos astronómicos a partir de sus líneas espectrales. Nuevas observaciones en regiones de longitudes de onda antes no explorados estarán disponibles gracias a proyectos como el Atacama Large Millimeter Array (ALMA). Se propone la extracción de descriptores estadísticos de las líneas de emisión y absorción de estructuras astronómicas con el fin de detectar patrones en líneas espectrales tanto de líneas individuales como conjuntos de líneas que correspondan a las mismas moléculas.

**Palabras Claves:** líneas espectrales: líneas de emisión, líneas de absorción; técnica: espectroscopía; método: análisis de datos.

# 1. Resumen Ejecutivo

El contexto de este informe es el de desarrollar una propuesta en el área de aprendizaje de máquina y minería de datos, desarrollando un proyecto colaborativo para el proyecto de Observatorio Virtual Chileno (ChiVO).

Se utilizarán datos provenientes del proyecto ALMA, con los cuales se aplicarán técnicas de análisis de datos para la detección, caracterización y clasificación de líneas espectrales.

En este documento se detalla la proposición de una herramienta de detección de líneas espectrales bajo la perspectiva de métricas y la búsqueda de patrones entre líneas de emisión y absorción tanto aisladas como en conjuntos de líneas pertenecientes a las mismas moléculas.

# Índice

## 2. Metodología de Trabajo

Para la proposición del problema se efectuaron una serie de reuniones tanto administrativas, con todas las universidades involucradas, como técnicas, con las universidades encargadas del área de minería de datos del proyecto ChiVO.

Con la ayuda de astrónomos fue posible realizar una propuesta que satisficiese los requerimientos de un problema astronómico utilizando técnicas de análisis de datos y aprendizaje de máquina.

Las reuniones llevadas a cabo con los proyectos de minería de datos permitieron dilucidar la problemática abordada en el proyecto. El flujo de reuniones se detalla a continuación:

| | |
|---|---|
| 29-08-2013 | Se propone abordar el tema de espectroscopía molecular y se compromete la búsqueda de potenciales problemas a resolver. |
| 06-09-2013 | Se expone la relación entre la naturaleza de los datos de ALMA y la espectroscopía molecular. Se compromete el estudio del estado del arte en identificación de líneas espectrales. |
| 27-09-2013 | Se describen las técnicas convencionales de espectroscopía para la identificación de líneas espectrales y las herramientas con las que se cuenta actualmente. |
| 08-10-2013 | Se proponen áreas de estudio iniciales para abordar el problema; se determina el problema a desarrollar como identificación de líneas espectrales a partir de descriptores estadísticos. |

Cuadro 1: Reuniones Técnicas del Proyecto

Por otra parte, el progreso del desarrollo del problema se apoya en la realización de reuniones mensuales con todas las universidades del proyecto para analizar avances y la entrega de hitos con el cumplimiento secuencial de los diferentes requerimientos del proyecto de Observatorio Virtual.

Se definieron estándares para todos los proyectos, con el fin de entregar un producto final consecuente con la idea de generar un paquete de herramientas astronómicas para ChIVO disponible para la comunidad astronómica. Se definió el uso del lenguaje de programación Python por su fácil integración con el software CASA.

# 3. Definición del Problema

La astronomía está enfrentando nuevos desafíos en como analizar grandes volúmenes de datos y como buscar o predecir eventos/patrones de interés. Un desafío sin precedentes para los radio-astrónomos hoy en día es desarrollar una máquina capaz de detectar, caracterizar y clasificar líneas espectrales provenientes de objetos astronómicos.

Dada la inmensa cantidad de datos siendo producidos en nuevas regiones de observación provistas por el Atacama Pathfinder Experiment (APEX), y por futuras misiones como el Stratospheric Observatory For Infrared Astronomy (SOFIA), el observatorio espacial HERSCHEL y el Atacama Large Millimeter Array (ALMA)[MSSW05], es fundamental contar con una herramienta que proporcione una forma automática de identificación de líneas de emisión y absorción.

Además, al identificar líneas en espectrógrafos en diferentes regiones de frecuencia y asociar cada línea a una frecuencia específica para cada átomo o molécula determinada, es posible determinar la composición quiímica de estructuras astronómicas [MSSW05].

## 3.1. Descripción

En este proyecto, nuestro objetivo es desarrollar un algoritmo de detección y clasificación automática de líneas espectrales de estructuras astronómicas pertenecientes a las observaciones proporcionadas por ALMA. El proceso de creación de un sistema de detección y clasificación automática comienza con la creación de descriptores estadísticos para líneas de emisión a través de su comportamiento a lo largo de diferentes longitudes de onda, así como la implementación de un modelo de aprendizaje de máquina para aprender como identificar y asignar cada frecuencia que presente una línea espectral a una molécula determinada en un diferente estado de energía.

# 4. Estado del Arte

El área de Machine learning tiene mucho potencial para ser aplicado en el área de la radio-astronomía y la identificación de líneas de emisión y absorción. Actualmente, la determinación de líneas espectrales se ha limitado al análisis a mano de datos para encontrar parámetros moleculares que permitan asociar los peaks en las mediciones de los espectrogramas a moléculas o átomos en ciertos estados de energía. Es evidente la falta de escalabilidad y lo impráctico que resulta este método manual para grandes cantidades de datos [PSP01]. Lo anterior es también complejizado por la dificultad a la hora de predecir nuevas coincidencias entre frecuencias y moléculas dada por las superposiciones de líneas. Estas mezclas de líneas (o blending) y superposiciones (o beams) son producto del ruido, así como la falta de sensibilidad para distinguir entre dos líneas en frecuencias cercanas. Lo anterior también puede producir peaks dobles en ciertas líneas [CTF$^+$13]. Un problema importante a la hora de identificar frecuencias subyace en líneas ópticamente delgadas, que tienden a dar resultados incorrectos. Usualmente, el uso de líneas de isótropos para su corrección resulta en un proceso costoso en tiempo e impractico para datos masivos [PSP01].

Nummelin et al. [NS98] propone el uso de un ajuste manual de las líneas a una forma arbitraria dada por una gaussiana, obteniendo por cada línea su frecuencia observada, el peak en el brillo de temperatura y el ancho de la velocidad (ancho total a media altura), para así proceder con la identificación de la línea al asociarla con una molécula en cierto estado de energía. Para la identificación de líneas, es necesario asumir temperatura y origen homogéneo, dado que la diferencia de temperatura cambia la relación en serie de intensidades de las líneas hiperfinas [NS00]. Es posible detectar patrones en las líneas que corresponden a la misma molécula a partir de intensidad relativa considerando que existe una razón entre diferencias de velocidad que es constante para un conjunto de líneas de emisión. Esto permite buscar patrones no tan solo de manera individual, sino que a través del análisis manual de series de líneas que se asocian a una misma molécula o átomo en sus diferentes estados energéticos.

El uso de ajustes de modelos de regresión y de indicadores de autocorrelación es una aproximación al problema que ha dado buenos resultados en diferentes áreas de la astronomía. La extracción de descriptores estadísticos de las series de tiempo ha permitido la clasificación de objetos estelares [DSA$^+$07, RSB$^+$11, SL13, KPB$^+$11]. Técnicas comunes para la clasificación en Machine Learning son la mezcla de Gaussianas [AND77], los árboles de decisión [Qui93], naive Bayes [DH73], Redes Neuronales [**?**], Support Vector Machines [CV95], Random Forest [**?**], entre otras. Estos métodos son modelos de análisis de datos que aprenden a predecir variables categóricas a partir de un set de otras variables de cualquier tipo. En nuestro caso, estos modelos buscarán predecir la molécula a la cual pertenece una línea o un conjunto de estas dadas las características de descritas a partir de su representación en descriptores estadísticos.

# 5. Planificación de Trabajo

## 5.1. Alcance del proyecto

### 5.1.1. Objetivo General

Desarrollar un algoritmo de detección y clasificación automática de líneas espectrales.

### 5.1.2. Objetivos Específicos

1. Determinar y caracterizar líneas espectrales a través de descriptores estadísticos de las representaciones numéricas de líneas de emisión.
2. Clasificar líneas espectrales o conjuntos de estas utilizando el catálogo Splatalogue.
3. Buscar patrones tanto a nivel individual como en conjuntos de líneas pertenecientes a las mismas moléculas.
4. Escoger e implementar un modelo de Aprendizaje de Máquina para la detección y clasificación de líneas espectrales no catalogadas.

## 5.2. Carta Gantt

| Inicio | Termino | Actividad |
|---|---|---|
| 06-09-2013 | 31-10-2013 | Formulación del Proyecto |
| 01-11-2013 | 15-11-2013 | Extracción de descriptores estadísticos de líneas espectrales conocidas. |
| 16-11-2013 | 30-11-2013 | Implementación de algoritmos de detección de líneas espectrales. |
| 1-12-2013 | 31-12-2013 | Búsqueda de patrones en líneas individuales |
| 1-01-2013 | 31-01-2013 | Búsqueda de patrones en conjunto de líneas |
| 01-03-2013 | 31-07-2013 | Implementación de algoritmo de clasificación de líneas |
| 01-08-2013 | 31-10-2013 | Integración del Proyecto en ChiVO |
| 1-11-2014 | 31-12-2014 | Validación del Proyecto |

Cuadro 2: Carta Gantt del Proyecto

## 5.3. Explicación hitos importantes

**Formulación del Proyecto** Definición de la metodología, el problema, los objetivos y la planificación del problema.

**Descripción numérica y búsqueda de patrones en líneas espectroscópicas** Detección y caracterización de líneas de emisión y absorción.

**Clasificación de líneas espectroscópicas** Análisis y resultados para selección de modelos de clasificación de líneas.

**Integración del Proyecto en ChiVO** Desarrollo de interfaces de integración del proyecto al conjunto de herramientas desarrolladas para ChiVO.

**Validación del Proyecto** Implementación de herramienta de forma pública en ChiVO y etapa de pruebas del software desarrollado.

## Referencias


[AND77] Dempster A., Laird N., and Rubin D. Maximum likelihood from incomplete data via the {EM} algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[Bre01] L Breiman. Random forests. *Machine learning*, pages 5–32, 2001.

[CTF+13] J. Cernicharo, B. Tercero, A. Fuente, J. L. Domenech, M. Cueto, E. Carrasco, V. J. Herrero, I. Tanarro, N. Marcelino, E. Roueff, M. Gerin, and J. Pearson. Detection of the Ammonium Ion in Space. *The Astrophysical Journal*, 2013.

[CV95] C Cortes and V Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, 1995.

[DH73] R Duda and P Hart. *Pattern Classification and Scene Analysis*. John Willey & Sons, 1973.

[DSA+07] J Debosscher, L Sarro, C Aerts, J Cuypers, B Vandenbussche, R Garrido, and E Solano. Automated supervised classification of variable stars. I. Methodology. *Astronomy and Astrophysics*, 475, 2007.

[KPB+11] Dae-won Kim, Pavlos Protopapas, Yong-ik Byun, Charles Alcock, Roni Khardon, and Markos Trichas. Qso selection algorithm using time variability and machine learning: selection of 1,620 qso candidates from macho lmc database. 2011.

[MSSW05] Holger S.P. Müller, Frank Schlöder, Jürgen Stutzki, and Gisbert Winnewisser. The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *Journal of Molecular Structure*, 2005.

[NS98] Ohishi M Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and Saito S. A Three-Position Spectral Line Survey of Sagittarius B2 between 218 and 263 GHz. I. The Observational Data. *The Astrophysical Journal Supplement Series*, 1998.

[NS00] Ohishi M Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and Saito S. A Three-Position Spectral Line Survey of Sagittarius B2 between 218 and 263 GHz. II. Data Analysis. *The Astrophysical Journal Supplement Series*, 2000.

[PSP01] T. R. Hunter D. C. Lis P. Schilke, J. Beneford and T. G. Phillips. A line survey of orion-kl from 607 to 725 ghz p. *The Astrophysycal Journal Supplement Series*, 2001.

[Qui93] J Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[RSB+11] J. W. Richards, D. L. Starr, N. R. Butler, J.S. Bloom, J. M. Brewer, A Crellin-Quick, J Higgins, R Kennedy, and M Rischard. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *The Astrophysical Journal*, 733, 2011.

[SL13] López M Aerts C Sarro LM, Debosscher J. Automated supervised classification of variable stars. (9918), 2013.

[WKP10] Y Wang, R Khardon, and P Protopapas. Shift-Invariant Grouped Multi-task Learning for Gaussian Processes. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 418–434. Springer Berlin / Heidelberg, 2010.

[WKPA09] Gabriel Wachman, Roni Khardon, Pavlos Protopapas, and Charles Alcock. Kernels for Periodic Time Series Arising in Astronomy. In Wray Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5782 of *Lecture Notes in Computer Science*, pages 489–505. Springer Berlin / Heidelberg, 2009.