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

Identicación de Líneas Espectrales Utilizando Métodos Estadísticos

Andrés Riveros, Karim Pichara, Diego Mardones,
Mauricio Solar, Marcelo Mendoza Jorge Ibsen, Lars Nyman, Eduardo Vera, Guillermo Cabrera,
Paola Arellano, Paulina Troncoso, Ricardo Contreras, Neil Nagar, Victor Parada.

Santiago, 1 de septiembre de 2014

Resumen

La detección e identificación automática de líneas espectrales es un problema astronómico que aún no ha sido resuelto. El uso de la espectroscopía permite describir la composición química de objetos astronómicos a partir de sus líneas de emisión. Nuevas observaciones en regiones de longitudes de onda antes no exploradas estarán disponibles gracias a proyectos como el Atacama Large Millimeter Array (ALMA). Con el uso de datos simulados basados en las observaciones que se tendrán a partir de ALMA, se propone una identificación de líneas emisión con el fin de identificar las moléculas que componen los objetos astronómicos. Para esto se propone un modelo que evalúe las posibles combinaciones de moléculas que dan origen al espectro observado. A partir de este algoritmo, los astrónomos podrán obtener una distribución de probabilidad de las posibles combinaciones de moléculas que componen los objetos astronómicos.

Palabras Claves: líneas espectrales: líneas de emisión; técnica: espectroscopía; método: aprendizaje de máquina.

1. Resumen Ejecutivo

En el presente documento se detallará el prototipo de la identificación de líneas de emisión de manera automatizada, los enfoques posibles para resolver el problema y la técnica propuesta en este proyecto para solucionar este problema.

Este proyecto se desarrolla en el marco del desarrollo de un proyecto colaborativo de varias universidades chilenas para la iniciativa del Observatorio Virtual Chileno (ChiVO). ChiVO es una plataforma en línea que pondrá a disposición de los astrónomos acceso a las mediciones del radio-telescopio Atacama Large Millimeter Array (ALMA). Además, proporcionará una serie de herramientas con las cuales podrán procesar y obtener información especializada de dichas mediciones de ALMA.

El radio-telescopio ALMA proporcionará cubos de datos que corresponden a dos dimensiones espaciales y una tercera dimensión en un rango de longitud de onda. Esto significa que se cuenta con una matriz de la cual en par espacial se puede obtener un espectrograma.

La Pontificia Universidad Católica de Chile (PUC) participa con el desarrollo de una herramienta de identificación de líneas espectroscópicas utilizando técnicas de aprendizaje de máquina y minería de datos. Esta herramienta consiste en un algoritmo que utiliza métodos estadísticos para obtener una distribución de probabilidad de posibles moléculas que dan origen a los espectrogramas observados.

Dado que en la fase en la que se encuentra el proyecto no se cuenta con datos reales, se utilizarán datos simulados con las características de las futuras mediciones de ALMA. Para esto se utilizará el software de simulación ASYDO, desarrollado en conjunto con este proyecto como parte de las herramientas disponibles en el proyecto ChiVO.

Con este conjunto simulado de espectros se pretende identificar líneas espectroscópicas a partir de un conjunto de moléculas con las cuales fueron simuladas dichas líneas espectrales. Se espera poder recuperar este conjunto de moléculas solo con el espectrograma observado. El resultado será una distribución de pertenencia de estas líneas espectroscópicas a diferentes configuraciones de moléculas.

La particularidad del algoritmo de identificación propuesto radica en la incorporación del factor probabilista en la identificación de los espectros, al consistir el resultado del algoritmo en distribuciones de probabilidad en vez de un resultado determinista.

A lo largo del documento se detalla el estado del arte del problema, la metodología con la cual se espera resolver el problema, el enfoque con el que será abordado el problema, la especificación de requerimientos del algoritmo a desarrollar y, finalmente, se describirá el prototipo de la solución propuesta, con su diseño e implementación.

Índice

1. Resumen Ejecutivo **2**

2. Estado del Arte **6**

3. Metodología de Trabajo **8**

3.1. Identificación de problemas a resolver 8

3.2. Simulación de Datos 9

3.3. Especificación de Requerimientos 9

4. Prototipo de solución **11**

4.1. Solución Propuesta 11

4.2. Diseño del Prototipo 11

4.3. Implementación 12

4.3.1. Etapa de Detección 12

4.3.2. Etapa de Predicción 15

5. Conclusiones **21**

Bibliografía **22**

Índice de figuras

1. Cubo de datos de ALMA, con dos dimensiones espaciales y una dimensión de frecuencia. . 8
2. Espectrograma con líneas de emisión solapadas. 11
3. Se desea recuperar la lista de moléculas con la que se simuló el espectrograma. 12
4. Cálculo del parámetro de sensibilidad y ejemplo de su aplicación. 13
5. Determinación de máximos locales para detección de líneas potenciales de emisión. 13
6. Espectro residual al sustraer una gausiana ajustada en una potencial línea de emisión. . . 14
7. Conjunto de gausianas ajustadas sobre las potenciales líneas de emisión detectadas. . . . 15
8. Se buscan moléculas por frecuencia dentro del rango de la varianza. 16
9. Se buscan frecuencias teóricas por especie (molecula) en todo el espectrograma. 16
10. Cálculo de cociente entre líneas observadas y líneas teóricas. 17
11. Para todas las líneas detectadas se ordenan sus cuocientes. 19
12. Se determina un parámetro de corte visualmente. 19

Índice de cuadros

1. Moléculas con las cuales se generó una línea espectral de prueba 17
2. Moléculas y sus respectivos cocientes encontrados con el algoritmo 18
3. Matriz de Confusión de las Predicciones 20
4. Métricas de la Predicción 20

2. Estado del Arte

La determinación de líneas espectrales según el método tradicional se limitaba al análisis manual de datos para encontrar parámetros moleculares que permitan asociar los peaks en las mediciones de los espectrogramas a moléculas en ciertos estados de energía.

La falta de escalabilidad de técnicas que no automatizadas, y lo poco práctico que resultan dichos método para grandes cantidades de datos [PSP01], añadido a la dificultad a la hora de predecir nuevas coincidencias entre frecuencias y moléculas dada por las superposiciones de líneas, ha impulsado a los astrónomos a buscar la automatización de esta tarea.

El problema de mezclas de líneas y superposiciones son producto de tanto ruido 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 isotopos para su corrección resulta en un proceso costoso en tiempo y por lo mismo no es apto 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 gausiana, 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 considerando las relaciones entre brillo de temperatura en un mismo espectro, es necesario asumir temperatura y origen homogéneo, dado que la diferencia de temperatura cambia la relación en serie de intensidades de líneas hiper-finas [NS00].

Esto es importante a la hora de utilizar datos simulados con el fin de representar fielmente las características físicas de las estructuras a utilizar para entrenar, de modo que el modelo sea posteriormente aplicable sin mayores variaciones al utilizar datos reales de ALMA.

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.

Los esfuerzos para desarrollar una herramienta automática de detección de líneas actualmente se limitan a herramientas semi-automáticas que utilizan como base complejos modelos físicos y químicos para la clasificación de líneas. Dichos modelos son aplicados en solo una medición en un espectro determinado, por lo que no se consideran las correlaciones existentes entre distintas mediciones de espectros para un mismo objeto. [SRC11].

Estas herramientas hacen uso de catálogos que contienen información sobre líneas espectroscópicas de moléculas y sus frecuencias teóricas de laboratorio, las que están disponibles públicamente en catálogos como Splatalogue [RMK08, Rem10].

Las técnicas anteriormente descritas no son escalables al no ser procesos automatizados y depender de

análisis o ajustes manuales que con la inminente llegada de enormes cantidades de datos provenientes de instrumentos como ALMA, dejan de ser aplicables. Por esto es necesario buscar algoritmos de clasificación que deleguen la tarea de identificar y clasificar líneas espectrales.

3. Metodología de Trabajo

En la metodología de trabajo se detalla el proceso a través del cual el problema a tratar en este proyecto fue concebido, los obstáculos que ha presentado y los pasos a seguir para hallar la solución al problema.

Este proyecto surge de la potencial capacidad del aprendizaje de máquina y de la minería de datos de ser utilizados como herramienta para la astronomía. La dificultad de los problemas astronómicos complejos utilizando minería de datos radica en que estos involucren una gran cantidad de datos a ser procesados.

ChiVO pondrá a disposición herramientas que utilizarán datos de ALMA, por lo que la cantidad de datos de mediciones astronómicas a procesar crecerá exponencialmente. Es por esto que una herramienta que automatice el procesamiento de dichos datos y entregue información útil y especializada a astrónomos será de gran utilidad.

3.1. Identificación de problemas a resolver

En una etapa inicial, se efectuaron una serie de reuniones administrativas con el conjunto de universidades que colaboran en el proyecto ChiVO. En estas reuniones se buscaba establecer como abordar y en qué se centraría cada universidad. El objetivo era desarrollar herramientas para los astrónomos que fueran útiles e innovadoras y que utilizaran los datos que estarán disponibles a partir de ALMA.

Así surgió la idea de desarrollar herramientas que aplicaran técnicas de minería de datos. Con los astrónomos involucrados en el proyecto fue posible realizar una propuesta de un problema astronómico. Cada proyecto de las universidades involucradas resolvería dicho problema con un enfoque diferente.

La información que el radio-telescopio ALMA entregará consiste en cubos de datos con dos dimensiones espaciales y una dimensión de longitud de onda. En cada par espacial se mide el brillo de temperatura para un rango de longitudes de onda, o su equivalente en una frecuencia determinada.

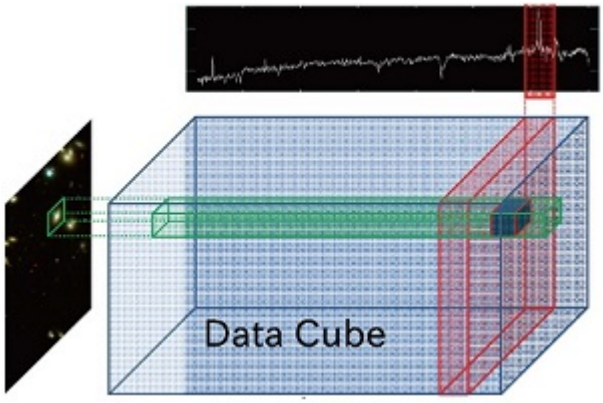

Figura 1: Cubo de datos de ALMA, con dos dimensiones espaciales y una dimensión de frecuencia.

Producto de la radiación que emiten los objetos estelares y de su composición, es posible observar líneas de emisión que son características para ciertos niveles de energía de las moléculas que conforman

el objeto astronómico.

Con esta información se contaría con una gran cantidad de líneas espectroscópicas, por lo que surge la idea de detectar dichas líneas espectrales como un problema astronómico de interés para aplicar técnicas de aprendizaje de máquina.

Por el lado del aprendizaje de máquina, se contará con suficientes datos para aplicar métodos estadísticos y desarrollar un algoritmo de identificación automática. Por parte de los astrónomos, se podrá contar con una herramienta que automatice la identificación de líneas de emisión en espectrogramas pertenecientes a objetos astronómicos.

Luego de definir el problema, era necesario determinar un set de datos con el cual comenzar a desarrollar el algoritmo. Dada la cantidad necesaria de espectros para desarrollar un modelo predictivo, y como actualmente no se cuenta con mediciones suficientes del radio-telescopio, se optó por utilizar una simulación de espectros.

3.2. Simulación de Datos

Para el desarrollo del algoritmo de identificación de líneas espectrales se utilizará el servicio web de datos simulados que proporcionará ChiVO. Este proyecto, llamado Astronomical SYntetic Data Observatory (ASYDO) se desarrolla en paralelo a este proyecto como parte de las herramientas que podrán utilizar los astrónomos como servicio web.

La simulación permitirá generar un set de entrenamiento para desarrollar el algoritmo de identificación propuesto en este proyecto. Al ser este servicio de simulación un input importante para el desarrollo del algoritmo, se han realizado una serie de reuniones para trabajar conjuntamente y obtener un set de datos simulados adecuado.

Con la ayuda de los astrónomos que participan en el proyecto ChiVO se ha determinado los parámetros necesarios para llevar a cabo la simulación. Se ha definido la complejidad y la importancia de replicar características determinadas que ayudarán a obtener curvas de espectros que se acerquen lo suficiente a los datos que se espera obtener a partir de las observaciones de ALMA.

Para que el algoritmo sea correctamente desarrollado con datos simulados, se deberá incluir en la meta-data de los cubos las líneas de emisión presentes en los espectros. Con esto será posible evaluar el modelo predictivo y determinar métricas para validar las predicciones.

3.3. Especificación de Requerimientos

El método de trabajo consiste en realizar 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 del observatorio virtual.

Se definieron estándares para todos los proyectos involucrados con el fin de entregar un producto final congruente con la idea de generar un paquete de herramientas astronómicas de ChIVO disponible para la

comunidad astronómica. Por lo mismo, se definió el uso del lenguaje de programación Python por su fácil integración con el software CASA como lenguaje común entre todos los servicios.

El algoritmo de detección de líneas espectroscópicas será un servicio web. Como input se recibirá un archivo de extensión .FITS, el cual contendrá la estructura de un objeto astronómico simulado con ASYDO. Como output se entregará un archivo con las líneas de emisión detectadas y la distribución de posibles moléculas a las que podrían pertenecer dichas líneas, con la probabilidad asociada a cada predicción realizada.

4. Prototipo de solución

4.1. Solución Propuesta

El algoritmo propuesto tiene como propósito la identificación de los componentes químicos que forman parte de diferentes objetos astronómicos a estudiar. Para esto, se analizan las líneas espectroscópicas a partir de los espectrogramas de dichos objetos astronómicos, con el fin de encontrar patrones y predecir su composición.

Para un objeto astronómico, observar sus líneas espectroscópicas puede ayudar a identificar las moleculas que lo componen, dado que para cada estado energético de dichas moléculas, al estar estas presentes en el objeto, se manifiestan líneas de emisión a lo largo de sus espectros observados en determinadas frecuencias.

El algoritmo toma ventaja de este comportamiento y busca patrones según la presencia de ciertas líneas. Cuando una molecula está presente en la composición de un objeto, probablemente deberían observarse una serie de líneas a lo largo del espectro, de tal forma que a menor cantidad líneas teóricas observadas de dicha molécula, menor será la probabilidad de que la molécula forme parte de la composición del objeto.

Así, si existe confusión a la hora de identificar una línea espectroscópica entre dos moléculas, es posible realizar una predicción probabilista de la molécula a la cual corresponde dicha línea basándose en la presencia de cada molécula a lo largo del espectro que se está midiendo.

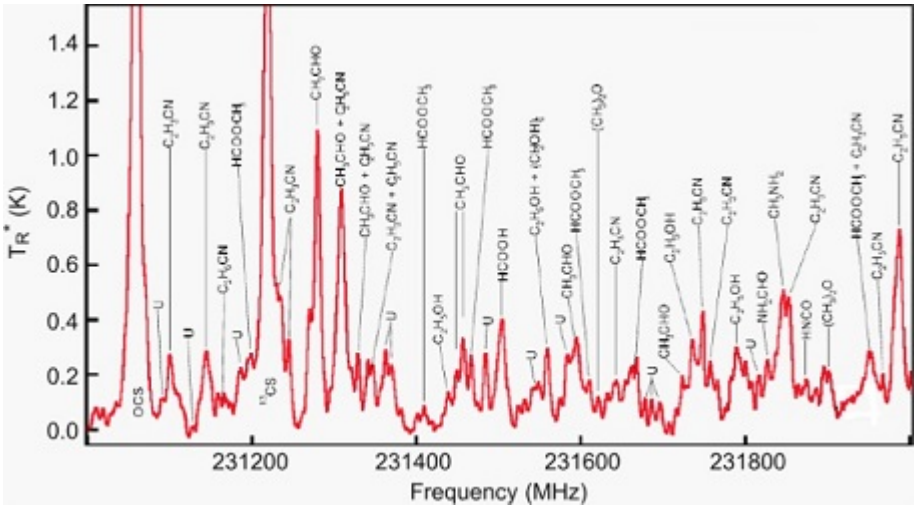

Figura 2: Espectrograma con líneas de emisión solapadas.

4.2. Diseño del Prototipo

Para el diseño del prototipo se utilizó el proyecto ASYDO [1]. La temperatura en estas simulaciones no posee unidades, ya que las magnitudes de las líneas son relativas a la línea más alta de CO, a la cual se le

[1] https://github.com/ChileanVirtualObservatory/ASYDO

ha asignado un valor arbitrariamente. Esto significa que los valores en sí mismos no poseen un significado, pero se respetan las diferencias relativas entre temperaturas.

El proceso para simular los cubos de prueba consistió en encontrar todas las moléculas y todos sus isotopos dentro de un rango de frecuencia. El rango de frecuencia utilizado para estas pruebas fue desde los 275 GHz hasta los 300 GHz, aproximadamente un cuarto de la banda número siete de ALMA. Se corrió un script de líneas combinadas de varios subconjuntos aleatorios de isotopos, subconjunto de tamaño variable del total de moléculas teóricamente existentes en esta ventana de frecuencias utilizada.

El objetivo del algoritmo es entonces recuperar la lista de moléculas con la cual se generaron los cubos de datos. Para realizar esta tarea solo se pueden utilizar los espectrogramas observados. A través de métodos estadísticos se pretende predecir la presencia de ciertas moléculas y validar dichas predicciones al conocerse las moléculas utilizadas para generar las simulaciones.

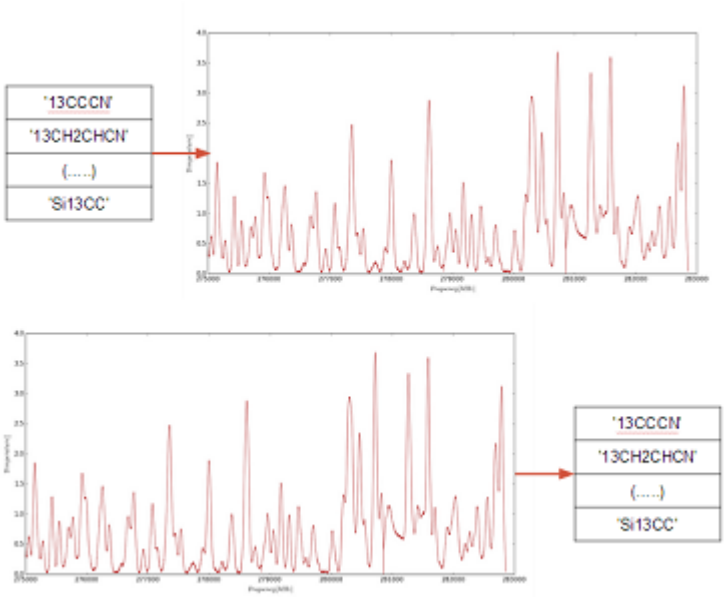

Figura 3: Se desea recuperar la lista de moléculas con la que se simuló el espectrograma.

4.3. Implementación

4.3.1. Etapa de Detección

El proceso de detección de líneas utiliza un parámetro de sensibilidad que determina si una medición es considerada una potencial línea espectroscópica. Esta sensibilidad depende de la desviación estándar del ruido en una región sin líneas visibles. Para obtener este parámetro se simuló un cubo de datos sin moléculas y se calculó la desviación estándar.

$$Threshold = 3 \cdot \sigma_{noise}$$

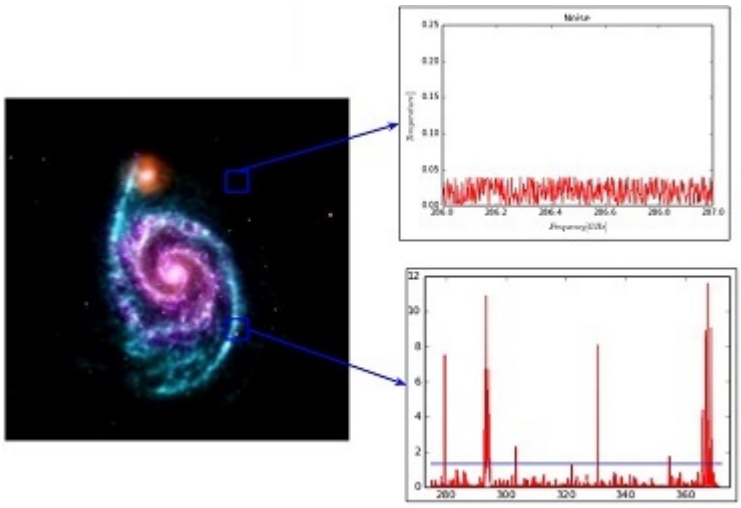

Figura 4: Cálculo del parámetro de sensibilidad y ejemplo de su aplicación.

A continuación, para cada punto espacial del cubo de datos, se toma cada espectrograma de manera independiente.

Se comienza con la determinación de los puntos máximos y mínimos de la curva observada. Es posible asignar un parámetro que determina la distancia mínima que debe existir entre máximos y mínimos para ser identificados. Sin este parámetro, la curva tendría una serie de falsos máximos y mínimos producto del ruido existente en las mediciones.

Cada máximo local detectado que está por sobre el parámetro de sensibilidad es un candidato a línea de emisión.

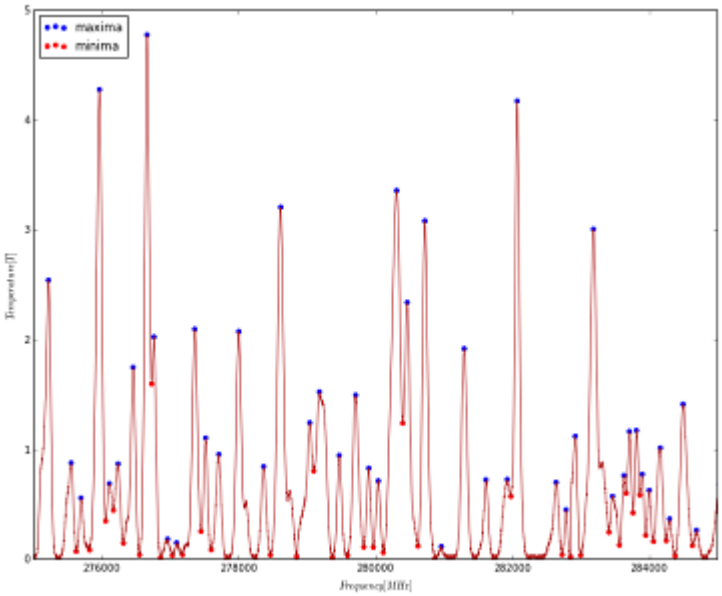

Figura 5: Determinación de máximos locales para detección de líneas potenciales de emisión.

A continuación, de manera iterativa, el algoritmo selecciona el peak más alto de la curva (máximo global). Una vez seleccionada la línea de emisión candidata, se ajusta una gausiana a esta y los parámetros de dicha gausiana (media y varianza) son guardados. Finalmente, cada gausiana ajustada es restada de la señal observada, con el fin de detectar líneas superpuestas que no son fácilmente observadas producto de que están demasiado cerca de líneas de de mayor magnitud.

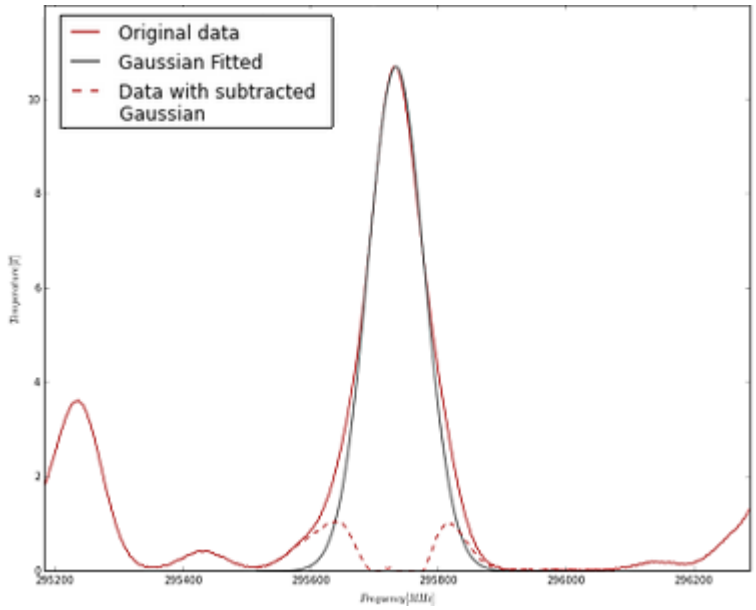

Figura 6: Espectro residual al sustraer una gausiana ajustada en una potencial línea de emisión.

Al final del proceso, el algoritmo entrega una lista de líneas candidatas con sus respectivas medias y desviaciones estándar. En la siguiente figura se puede ver la lista de gausianas ajustadas en el espacio de medición simulado.

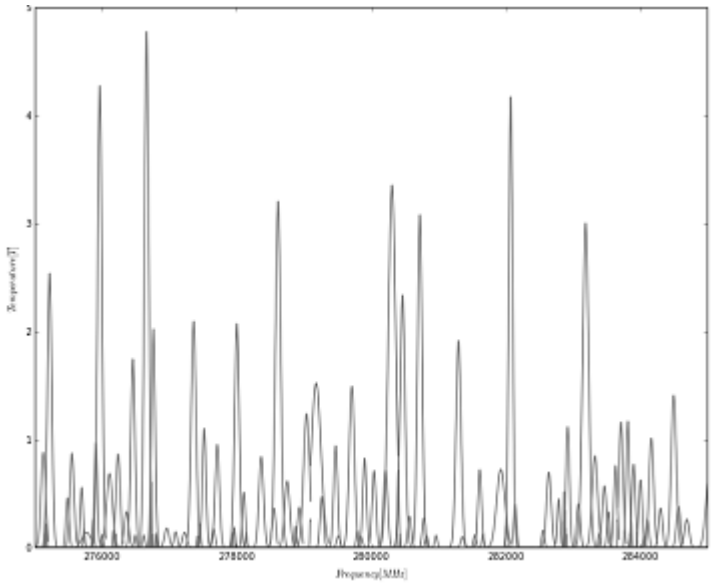

Figura 7: Conjunto de gausianas ajustadas sobre las potenciales líneas de emisión detectadas.

4.3.2. Etapa de Predicción

La siguiente etapa es la predicción. La formulación para el algoritmo es la siguiente:

Sea $\mathcal{L} = \{l_1, ..., l_n\}$ el conjunto de todas las frecuencias de las líneas observadas.

Sea $\mathcal{I} = \{i_1, ..., i_m\}$ el conjunto de todos los isotopos existentes en el rango de estudio.

Un modelo se define como $\mathcal{F} = \{F : \forall l \in \mathcal{L}, \exists i \in \mathcal{I} \mid F(l) = i\}$

Algoritmo:

1. $\delta\mathcal{L} = \{\delta l_1, ..., \delta l_n\}$ es el conjunto de rangos de búsqueda, donde cada elemento corresponde a cada rango de una línea detectada.

2. Para una línea detectada, sea su rango de búsqueda $\delta l_u \in \delta\mathcal{L}$, $u \in \{1, ..., n\}$

   Sea $\mathcal{I} = \{f_{i_1}, ..., f_{i_m}\}$ el conjunto de todos los isotopos teóricos que existen en $\delta\mathcal{L}$

   Cada isotopo teóricamente posible $i_v \in \mathcal{I}$ cumple $f_{i_v} \in \delta\mathcal{L}$. Por lo que un modelo posible $F_v^{l_i}$ asume que $F(l_i) = i_v$, $v \in \{1, ..., m\}$

   Se define el cociente $cuo_{u,v} = \frac{|F \cap F_v^{l_i}|}{|F|}$

3. Finalmente, para escoger el modelo $F_v^{l_i}$ que mejor describe al espectrograma observado,se busca $F_v^{l_i} = argmax_v\{cuo_{u,v}\}$

Puesto en palabras, para cada gausiana ajustada en el paso de detección, el algoritmo predice a que isotopo dicha línea detectada pertenece, y para esto, se utilizan dos criterios:

Se utiliza la varianza de la gausiana para determinar un rango de búsqueda. En este rango de búsqueda se obtienen todas las moléculas a las que la línea que se está evaluando puede pertenecer.

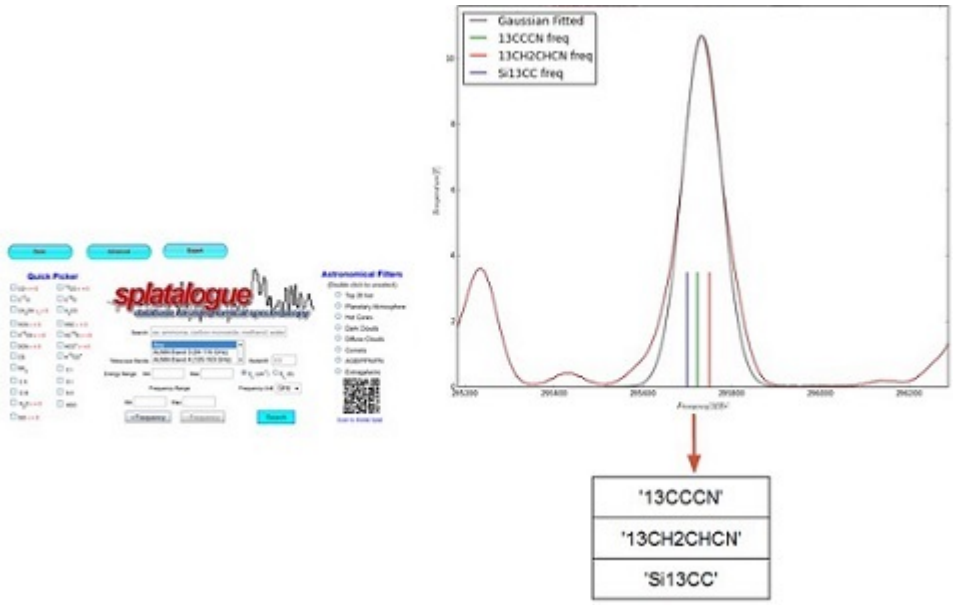

Figura 8: Se buscan moléculas por frecuencia dentro del rango de la varianza.

A continuación, se utiliza el catalogo de líneas espectroscópicas teóricas Splatalogue para obtener la lista de todas las frecuencias teóricas que existen para las moléculas encontradas en la etapa anterior.

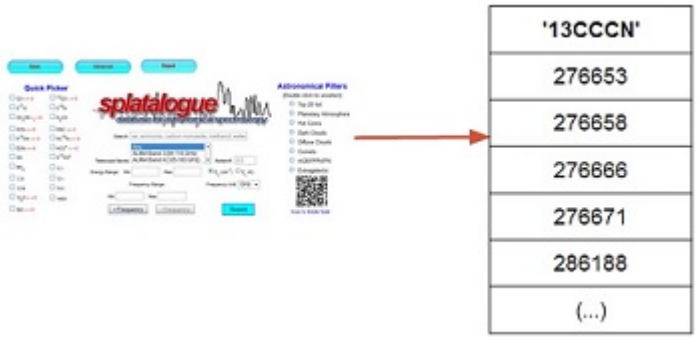

Figura 9: Se buscan frecuencias teóricas por especie (molecula) en todo el espectrograma.

Luego, el algoritmo calcula un indicador heurístico de pertenencia, el cual consiste en:

- Se cuenta el número de líneas teóricas presentes para un isotopo en particular en el rango de todas las gausianas detectadas.

- El número total de líneas que debería estar presente si la gausiana que se está evaluando se instanciara como un isotopo en particular.

Así, para cada gausiana se calcula el cociente de líneas observadas sobre líneas teóricas totales.

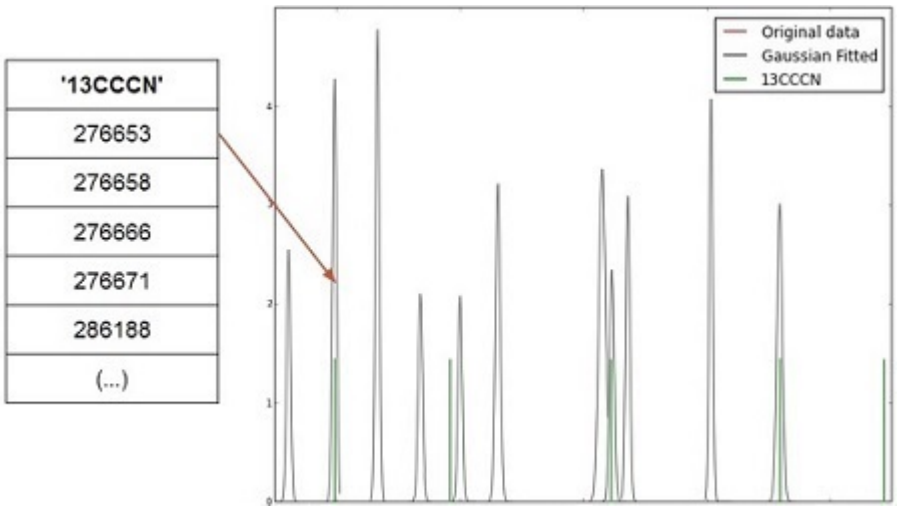

Figura 10: Cálculo de cociente entre líneas observadas y líneas teóricas.

$$\frac{NumberofLinesPresent}{NumberofLinesShouldBeThere}$$

Por ejemplo, para una simulación que tenga una línea simulada con las siguientes líneas presentes:

Molécula	Frecuencia	Temperatura
13CCCN	295.741	1.2261414597
13CCCN	295.736	1.22610908999
13CCCN	295.723	1.22603385007
13CCCN	295.727	1.22588330236
Si13CC	295.700	1.00924946246
Si13CC	295.763	0.530384205761

Cuadro 1: Moléculas con las cuales se generó una línea espectral de prueba

El algoritmo encuentra las siguientes moléculas y calcula un cociente para cada una:

Molécula	Líneas Observadas	Frecuencias Teóricas	Cociente
13CCCN	36	36	1.0
Si13CC	36	38	0.947368421053
CH3CCD	22	33	0.666666666667
c-H2C3O	58	102	0.56862745098
CH3CH213CN	284	521	0.545105566219
13CH2CHCN	180	340	0.529411764706
CH3OHvt=0	37	74	0.5
HNCv2=1	1	2	0.5
CH2CHCNv=0	170	347	0.489913544669
CH313CH2CN	279	570	0.489473684211
c-HCC13CH	21	43	0.488372093023
CH3CH2CNv=0	327	679	0.481590574374
g-Ga-(CH2OH)2	720	1499	0.480320213476
CH213CHCN	147	311	0.472668810289
H2CCCHCN	302	657	0.459665144597
H2CCNH	13	29	0.448275862069
c-H2COCH2	91	220	0.413636363636
t-HCOOH	32	83	0.385542168675
a-H2CCHOH	50	132	0.378787878788
s-H2CCHOH	36	100	0.36
HDS	2	8	0.25

Cuadro 2: Moléculas y sus respectivos cocientes encontrados con el algoritmo

Como se puede observar, las dos primeras moléculas observadas corresponden efectivamente a las moléculas que se utilizaron para simular dicha línea. Así, el siguiente paso es determinar un parámetro de corte que separe los cocientes de las líneas presentes en la simulación de los que no lo están.

Al correr el algoritmo para cada gausiana ajustada encontrada en la etapa de detección, y ordenar de mayor a menor los cocientes calculados para cada una de estas, se obtiene el siguiente gráfico, donde los puntos azules corresponden a cocientes de moléculas presentes en la gausiana correspondiente, y los puntos rojos a moléculas no presentes.

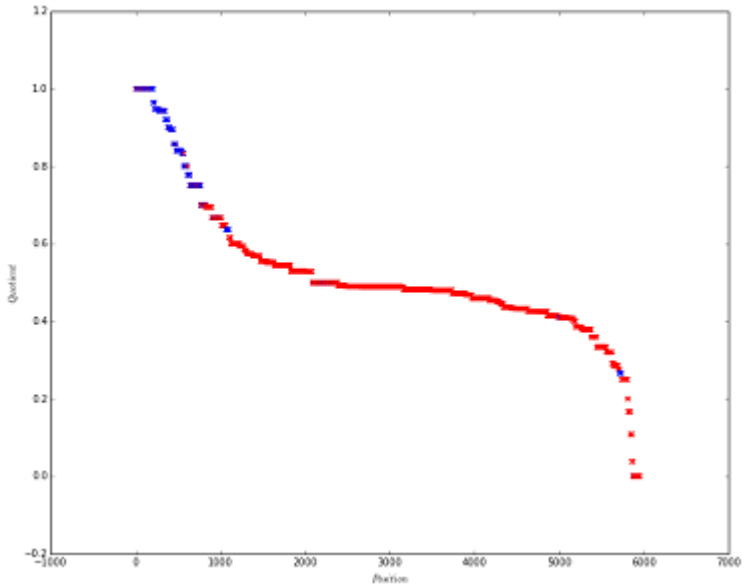

Figura 11: Para todas las líneas detectadas se ordenan sus cuocientes.

Para dar una idea del rendimiento actual de esta heurística, se procedió a determinar visualmente un parámetro de corte.

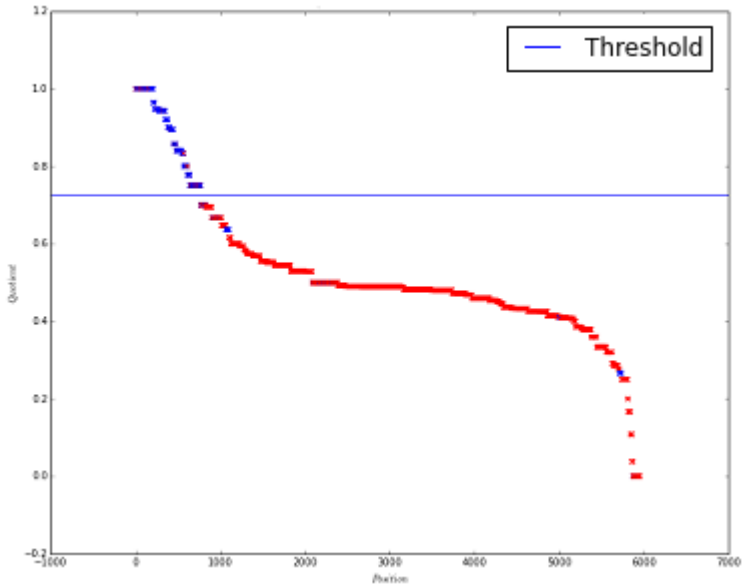

Figura 12: Se determina un parámetro de corte visualmente.

Así, fue posible obtener las siguientes métricas de validación de las predicciones[HK01, HTF13]:

	Presentes	No Presentes
Presentes	632	171
No Presentes	136	5014

Cuadro 3: Matriz de Confusión de las Predicciones

	Presentes (%)	No Presentes (%)
Precision	82.29	97.36
Recall	78.71	96.70
F-Score	80.46	97.03

Cuadro 4: Métricas de la Predicción

Como se puede observar, la predicción de moleculas no presentes en una línea tiene un porcentaje muy elevado, sin embardo, asegurar que una línea está compuesta de cierta molécula tiene un porcentaje que aún es mejorable de probabilidad de estar correcto.

5. Conclusiones

La métrica de heurística para determinar la presencia de moléculas en la composición de objetos astronómicos es un primer paso para dar con la solución final. Una extensión natural del algoritmo presentado es la búsqueda de un parámetro de separación adecuado que puede ser determinado por un algoritmo de regresión entrenado con una serie de cubos evaluados con el prototipo expuesto.

El siguiente paso consiste en la formalización de la etapa de predicción. Esto significa reemplazar la medida heurística por un enfoque Bayesiano que permita encontrar dependencias entre líneas y asignar importancias relativas a las distintas presencias de isotopos.

Por otra parte, aún es necesario incorporar la información de distancia entre una línea teórica y una línea observada. Esto se puede implementar con un conteo que incorpore pesos que favorezcan a las frecuencias teóricas más cercanas a la observación. Así, se puede definir una distancia exponencial que para una línea teórica que coincida con la media de la gausiana ajustada sumará uno al conteo, y para una línea caiga en el límite del rango de búsqueda, se contará cero.

Por otra parte, la incorporación de una probabilidad preferencial que asigne más importancia al conteo de líneas con mayores magnitudes de temperatura ayudará a restar el efecto del ruido, al no considerar presencias que estén muy cerca del parámetro de sensibilidad en la etapa de detección.

Para resolver el solapamiento de líneas, se evaluará el uso de Dirichet process mixture con el fin de encontrar la mejor combinación de moléculas para identificar el origen de ciertas líneas, pero dada la magnitud del problema, es necesario realizar la etapa de predicción para acotar el espacio de búsquedas e iterar en la combinación de moléculas posibles en dicho espacio más manejable.

Finalmente, la elaboración del modelo probabilista con un enfoque Bayesiano debe considerar la información disponible a través de la asignación de Priors adecuados, para así determinar a través de una función fitness qué combinación de moléculas genera de mejor manera el espectro observado.

Referencias

[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.

[HK01] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, USA, 2001.

[HTF13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Science, USA, 2013.

[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.

[Rem10] A. J. Remijan. Splatalogue - Motivation, Current Status, Future Plans. 2010.

[RMK08] A. J. Remijan and A Markwick-Kemper. Splatalogue: Database for Astronomical Spectroscopy. 2008.

[SRC11] Peter Schilke, Rainer Rolffs, and Claudia Comito. Analysis tools for spectral surveys. *Proceedings of the International Astronomical Union*, 7:440–448, 6 2011.