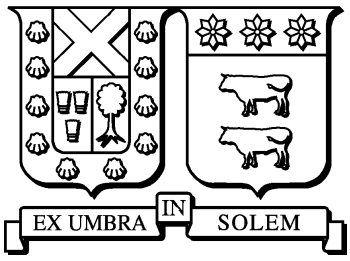
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"IDENTIFICACIÓN DE LÍNEAS ESPECTRALES UTILIZANDO MODELOS DE MEZCLAS PARA LA RADIOASTRONOMÍA"

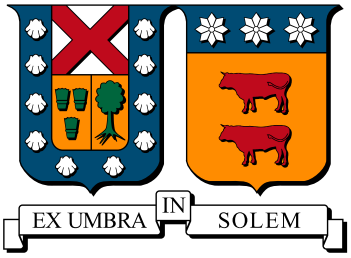
SEBASTIÁN IGNACIO ARANDA SÁNCHEZ

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: MARCELO MENDOZA

DECIEMBRE 2019

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"IDENTIFICACIÓN DE LÍNEAS ESPECTRALES UTILIZANDO MODELOS DE MEZCLAS PARA LA RADIOASTRONOMÍA"

SEBASTIÁN IGNACIO ARANDA SÁNCHEZ

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: MARCELO MENDOZA
PROFESOR CORREFERENTE: MAURICIO SOLAR

DECIEMBRE 2019

Agradecimientos

Esta memoria ofrece una herramienta de análisis de datos de astronomía que surge desde el observatorio virtual chileno ChiVO, proyecto FONDEF IT 15110041, un proyecto colaborativo entre universidades chilenas.

Para evaluar el algoritmo se hace uso de los siguientes datos de ALMA:
ADS/JAO.ALMA#2011.0.00001.SV, ADS/JAO.ALMA#2011.0.00009.SV,
ADS/JAO.ALMA#2012.1.00275.S, ADS/JAO.ALMA#2012.1.00346.S,
ADS/JAO.ALMA#2012.1.00395.S, ADS/JAO.ALMA#2013.1.00233.S y
ADS/JAO.ALMA#2013.1.01268.S.

Resumen

La astroinformática ha tomado mucho impulso debido a la construcción de telescopios cada vez más grandes y sofisticados, como el Atacama Large Millimeter/submillimeter Array. Estos instrumentos poseen una mejor resolución espectral que ha establecido un nuevo desafío en la forma en que los datos de astronomía son analizados. En particular, los cubos de datos producidos por proyectos de radioastronomía han generado una explosión en el volumen de datos capturado, esto hace más complejas algunas tareas como el análisis de líneas espectrales. Por esta razón es fundamental el desarrollo de herramientas de software que permitan realizar análisis automático de los datos.

Este trabajo propone un novedoso método para clasificar líneas espectrales utilizando modelos de mezclas. La propuesta esta basada en un algoritmo llamado Labeled Latent Dirichlet Allocation, un modelo generativo probabilístico capaz de describir documentos como mezclas de palabras sobre tópicos. Aquí cada espectro es representado como una mezcla de transiciones sobre especies moleculares. Una base de datos de líneas espectrales llamada Splatalogue es usada para entrenar diferentes modelos basados en el tipo de objeto y banda de frecuencia. El algoritmo es evaluado utilizando el modelo para analizar surveys de líneas espectrales y cubos de datos reales producidos por observaciones de ALMA. El potencial de esta propuesta es la habilidad de modelar datos de alta dimensionalidad y luego utilizar inferencia posterior para clasificar nuevas observaciones espectrales. Los resultados muestran que el algoritmo puede ser utilizado como herramienta para clasificar líneas espectrales individuales en cubos de datos con hasta un 97 % de accuracy.

Palabras clave. radioastronomía, ALMA, clasificación de líneas espectrales, labeled latent dirichlet allocation, modelos de mezclas.

Abstract

The discipline of astroinformatics has grown a lot over the past few years thanks to the creation of bigger and more sophisticated telescopes, such as the Atacama Large Millimeter/submillimeter Array. With better spectral resolution in data, a new challenge is set in the way astronomical data is analyzed. In particular, data cubes produced by radioastronomy projects have generated an explosion in the volume of data retrieved. Some tasks, such as the identification of spectral lines becomes more complex. For this reason it is essential to develope accurate analysis tools that allows data to be processed automatically.

This works propose a novel method in the way spectra can be classified. The approach is based on an algorithm used in the world of Text Mining, named Latent Dirichlet Allocation, a probabilistic generative model capable of describing documents as a random mixture of words over topics. Here, each spectrum is represented as a mixture of transitions over species. A spectral line transitions database named Splatalogue is used to train different models based on the type of observed object or ALMA band. The algorithm is evaluated using the model to analyze real world data cubes and spectral line surveys from radioastronomy observations of ALMA. The main advantage of the proposal is the ability to model sparse and high dimensional data using posterior inference to classify new spectral observations. Results shows that L-LDA can be used to clasifiy spectral lines on data cubes with up to 97 % of accuracy.

Keywords. radioastronomy, ALMA, spectral line classification, labeled latent dirichlet allocation, topic modeling.

Índice de Contenidos

Glosario

- NRAO (National Radio Astronomy Observatory): es una entidad perteneciente a la Fundación Nacional para la Ciencia. Se encarga de proveer instalaciones de última tecnología para la radio astronomía.
- ALMA (Atacama Large Millimeter/submillimeter Array): radio observatorio instalado en el llano de Chajnantor al norte de Chile.
- VO (Virtual Observatory): termino ocupado para referirse a los observatorios virtuales, entidades científicas encargadas del estudio de herramientas para el procesamiento de datos astronómicos.
- FONDEF (Fondo de Fomento al Desarrollo Científico y Tecnológico): organización que promueve la vinculación entre instituciones de investigación, empresas y otras entidades con el objetivo de generar proyectos de investigación aplicada y de desarrollo tecnológico.
- ChiVO (Chilean Virtual Observatory): observatorio virtual chileno. Encargado del de desarrollo de herramientas de software para el procesamiento de datos de astronomía.
- FITS (Flexible Image Transport System): es el formato de archivo estándar con el cual se almacenan los datos que produce cada observatorio astronómico.
- ROI (Region of interest): ejemplos de un conjunto de datos que poseen información valiosa. En este contexto se refieren a zonas con alta intensidad de emisión.
- ML (Machine Learning): área de la inteligencia artificial que estudia el funcionamiento de algortimos capaces de actuar sin ser programados explícitamente. Estos algoritmos

se sustentan en modelos basados en datos.

- TM (Text Mining): sub área del Machine Learning que estudia algoritmos capaces de analizar y procesar grandes conjuntos de texto.
- NLP (Natural Language Processing): campo de ciencias que se encarga el procesamiento de lenguaje natural digital.
- L-LDA (Labeled Latent Dirichlet Allocation): modelo estadístico del conjunto de los Topic Models.

Capítulo 1

Introducción

La astroinformática es un campo que ha tomado mucho impulso en los últimos años gracias a la creación de proyectos astronómicos de escala mundial. En particular, la creación de observatorios como el Atacama Large Millimeter/Submillimeter Array (ALMA) [1] o el Large Synaptic Survey Telescope (LSST) [4] han generado nuevas necesidades relacionadas con la captura, almacenamiento, transporte, análisis y descubrimiento de datos. Se estima que ALMA va a generar 200 TB de datos observacionales brutos. La gran sensibilidad que poseen los nuevos observatorios, motivará a la comunidad astronómica a invertir tiempo de investigación en generar soluciones tecnológicas innovadoras. Al año 2011 se contaba con aproximadamente 1 *petabyte* (PB) de datos públicos disponibles. Las proyecciones indican que para el año 2020, más de 60 PB de datos estarán disponibles para realizar estudios [9]. Propuestas innovadoras relacionadas con el descubrimiento de datos serán de gran interés para los astrónomos, sin embargo, es necesario que estas técnicas respondan al crecimiento del volumen de datos generado.

Agrupaciones de carácter científico conocidas como Observatorios Virtuales (VO) [7] se encargan de proveer distintos servicios informáticos que permiten satisfacer las necesidades actuales de la astronomía. Los servicios provistos se enfocan en almacenamiento, transporte, análisis y descubrimiento de datos. La Universidad Tećnica Federico Santa María participa de la administración del VO chileno, Chilean Virtual Observatory (ChiVO) [2]. En ChiVO se desarrollan distintas técnicas orientadas al análisis y almacenamiento de datos astrnómicos.

Actualmente se trabaja en la clasificación de líneas espectrales y este trabajo de memoria busca ser una propuesta a esta tarea, desarrollando un algoritmo capaz de identificar líneas de emisión en cubos de datos de ALMA y *surveys* de líneas espectrales. Lo anterior se enmarca en el Fondo de Fomento al Desarrollo Científico y Tecnológico (FONDEF), proyecto FONDEF IT 15110041.

Los trabajos enfocados en la clasificación automática de lineas espectrales son recientes y las técnicas utilizadas hasta el momento no son escalables o se basan en modelos teóricos muy complejos. Las aproximaciones varían en el tipo de algoritmo aplicado para entrenar modelos y en el alance de los datos de testeo; Riveros (2016) [27] propone un algoritmo aplicando *Sparse Coding*, basándose en aprendizaje no supervisado; Miranda (2015) [23] propone una solución utilizando Reglas de Asociación; Barrientos (2016) [8] postula una forma clasificar lineas espectrales utilizando métodos del mundo de Máquinas de Aprendizaje (*Machine Learning*, ML), utiliza los algoritmos Máquinas de Soporte Vectorial (*Support Vector Machines*, SVM) y Redes Neuronales Artificiales (*Artificial Neural Networks*, ANN). Estas aproximaciones generan sus modelos sobre datos de observaciones sintéticos para evaluar los algoritmos. Finalmente uno de los trabajos más recientes, y sobre el cual se respalda esta propuesta corresponde al método propuesto por Mendoza (2017) [22], donde utilizan modelos de mezclas (*topic models*) para resolver el problema. Mendoza entrena distintos modelos de mezclas utilizando datos de Splatalogue, luego clasifica las especies capturando la coocurrencia de transiciones.

Esta propuesta no representa una solución definitiva al problema de identificación automática de líneas espectrales, sino más bien una herramienta de apoyo para los astrónomos interesados en analizar conjuntos de datos de gran tamaño, del era del Big Data en astronomía. La idea central consiste en modelar espectros de observaciones de radioastronomía, utilizando una herramienta proveniente del mundo de Minería de Texto (*Text Minning*, TM), denominada Labeled Latent Dirichlet Allocation (L-LDA) [26], una variación de Latent Dirichlet Allocation (LDA) [10]. Estas herramientas utilizan modelos generativos probabilístico para modelar la coocurrencia de elementos en una colección de datos. Estos modelos representan los datos como una mezcla finita sobre un conjunto latente variables, en otras palabras se

generan esquemas implícitos que son representados por distribuciones de probabilidad sobre los mismos datos. Una aplicación de estos modelos corresponde a la identificación de tópicos en colecciones de texto, donde cada tópico se representa como una mezcla de palabras. En este trabajo se busca modelar la existencia de líneas espectrales en observaciones de radioastronomía, donde las ventanas espectrales pueden llegar abarcar grandes rangos de frecuencia. Para esto se entrenan diversos modelos utilizando el algoritmo L-LDA. Los datos son obtenidos del catálogo de líneas espectrales Splatalogue [5] que contiene más de 5.4 millones de transiciones registradas. Luego se evalúa la capacidad de predicción del modelo con datos reales de observaciones registradas en *surveys* y en cubos de datos de ALMA. Los resultados indican que el algoritmo puede ser utilizado como herramienta para clasificación de especies moleculares sobre cubos de datos de ALMA, alcanzando un Accuracy del 98 %.

1.1. Objetivo Principales

Se quiere desarrollar un algoritmo capaz de capturar la presencia de lineas de emisión de distintos elementos químicos en cubos de datos producidos por ALMA y en conjuntos de transiciones de *surveys* de líneas espectrales.

1.2. Objetivos Específicos

- Extender y validar el trabajo realizado por M. Mendoza et. al. [22].
- Proponer una aplicación del algoritmo L-LDA como modelo para realizar clasificación de datos en astronomía.
- Evaluar las características influyentes en el poder de predicción del algoritmo clasificando líneas de emisión.
- Evaluar el catálogo de líneas espectrales Splatalogue, como conjunto de datos para entrenar clasificadores de líneas espectrales.

Capítulo 2

Marco Teórico

2.1. Espectroscopia

La espectroscopia es una técnica que permite analizar la interacción ente la materia y la radiación a través de la recepción de ondas electromagnéticas, energía emitida por la materia en estado de excitación.

Gracias a los avances de la física moderna, tenemos un modelo para describir el comportamiento de los átomos, partícula bloque de la materia; los átomos están compuestos de un núcleo sólido con carga positiva y de una nubes de electrones orbitando con carga negativa. Se plantea que los electrones se encuentran orbitando en ciertos niveles de energía discretos, así cuando un átomo se encuentra en estado de excitación sus electrones saltan entre estos niveles de energía emitiendo fotones. Los fotones son paquetes de energía discretos en forma de radiación electromagnética, esto se puede apreciar a un lado izquierdo de la figura 2.1.

Una onda electromagnética puede ser descrita en términos de longitud de onda λ y frecuencia f. Estas magnitudes se relacionan junto a la constante de la velocidad de la luz $c = 299792458\,[m/s]$ según la ecuación de onda:

$$c = \lambda f \tag{2.1}$$

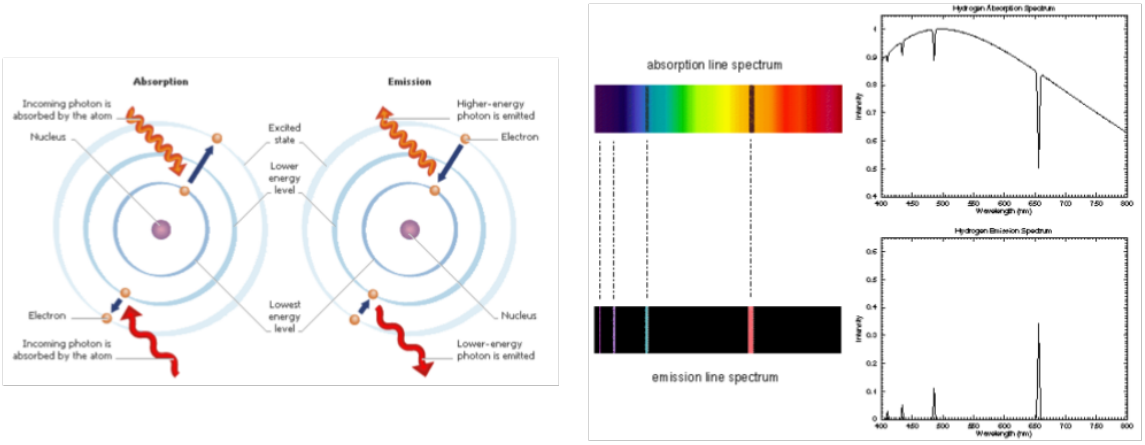


Figura 2.1: Emisión y absorción de fotones representado con el modelo atómico clásico a un lado izquierdo. Representación de líneas de emisión y absorción en un espectro a un lado derecho.

Además la energía E y frecuencia f de una onda electromagnética están relacionadas según la ecuación de Planck-Einstein:

$$E_t = hf \tag{2.2}$$

La ecuación 2.2 indica que la energía de una transición y su frecuencia son directamente proporcionales, donde $h = 6,626070040 \cdot 10^{-34}[J \cdot s]$ es la constante de Planck. Otra relación muy importante definida por Planck corresponde al modelo de cuerpo negro, un objeto hipotético que absorbe toda la radiación recibida. Este tipo de objetos se comportan como perfectos emisores de radiación electromagnética y para el análisis espectroscópico se asume que las fuentes de emisión se comportan de esta manera. Se establece que la distribución del espectro generado por la radiación de un cuerpo negro depende solo de su temperatura.

Otro factor físico que influye en una observación astronómica es el Efecto Doppler. Se define que la frecuencia observada f de una onda se relaciona con la frecuencia de reposo o frecuencia emitida f_0, y con la velocidad radial v_r de la fuente con respecto al observador

según la siguiente relación:

$$f = \left(\frac{c}{c + v_r}\right) f_o \tag{2.3}$$

En otras palabras, la frecuencia que se percibe de una onda emitida por un objeto en movimiento, difiere de la frecuencia original. Este efecto tiene gran influencia en los cálculos realizados en astronomía, debido a que todos los objetos que observamos se están acercando o alejando con respecto a un observador en la tierra. Dos fenómenos consecuentes se presentan a continuación:

- **Redshift**: conocido también como corrimiento al rojo, representa el grado de alejamiento de un objeto que emite radiación electromagnética. Esto se refleja en desplazamientos de las frecuencias hacia valores menores (también existe el Blueshift que representa un corrimiento hacia valores mayores de frecuencia), por esto es necesario calcular el nivel de desplazamiento para calcular la frecuencia emitida original del objeto. Para valores de $v_r \ll c$ el Redshift z se calcula como:

$$z = \frac{f_0 - f}{f} \approx \frac{v_r}{c} \tag{2.4}$$

- **Doppler Broadening**: conocido también como ensanchamiento de doppler, refleja el grado de excitación de las partículas de un objeto que emite radiación. Esto se produce debido a la distribución de velocidad de los átomos y moléculas del objeto. Por esta razón las líneas espectrales no se aprecian como una emisión en una única frecuencia sino como curvas con perfiles gaussianos (ver figura 2.2), con anchos variables en función de la temperatura y masa del objeto. Estos perfiles se define por la siguiente desviación estándar:

$$\sigma_f = \sqrt{\frac{8kT \ln 2}{mc^2}} f_0 \tag{2.5}$$

  Donde $k = 1,380648528 \cdot 10^{-23}[J \cdot K^{-}1]$ es la constante de Boltzman, T es la temperatura y m es la masa del objeto.

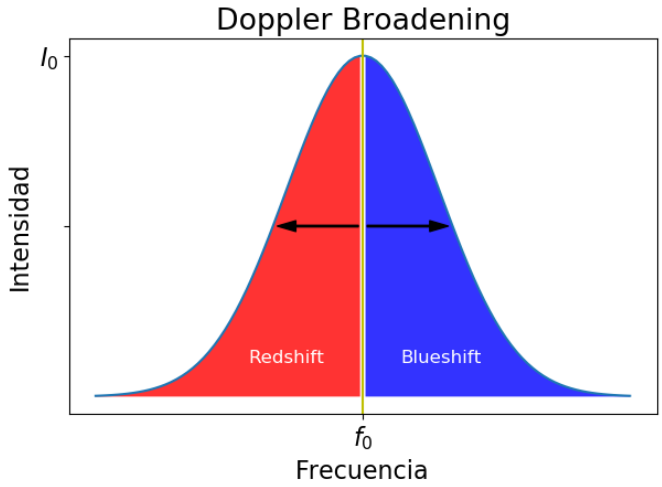


Figura 2.2: Esquema gráfico del efecto producido por el ensanchamiento de Doppler. Una línea espectral se observa como un pico de intensidad centrado la frecuencia específica de la transición.

Teniendo en consideración los efectos descritos es posible determinar diversas propiedades de los objetos que se estudian en astronomía. Una de estas propiedades es la composición molecular de los objetos de interés. Cada átomo emite radiación electromagnética en un patrón determinado por los cambios de nivel en las órbitas de los electrones; los diferentes patrones generados por diferentes moléculas se relejan en las observaciones realizadas, por esta razón es posible determinar la composición molecular en función de patrones ya estudiados. Los patrones descritos se muestran como curvas de intensidad centradas en una frecuencia particular [3]. Actualmente los espectros son almacenados de forma digital y son representados como gráficos de intensidad en función de la frecuencia o longitud de onda, una representación gráfica de esto se aprecia a un lado derecho de la figura 2.1.

Hasta el momento la tarea de catalogar lineas espectrales corresponde a un trabajo complejo y tedioso. Los astrónomos deben identificar y etiquetar las observaciones de manera manual. La identificación de especies en el espacio requiere la comparación directa de frecuencias de líneas de emisión o absorción con mediciones espectroscópicas realizadas en experimentos de laboratorio [11].

2.2. Catálogo de líneas espectrales

Cuando un científico publica el descubrimiento de una nueva especie molecular y los resultados son validados se procede a registrar la información en catálogos de líneas espectrales. Existen diversas agrupaciones científicas encargadas de mantener estos catálogos actualizados y de proveer servicios para que la comunidad astronómica aproveche estos datos.

La solución propuesta en este trabajo de memoria se apoya en un catálogo de lineas espectrales llamado Splatalogue, creado por la organización NRAO (National Radio Astronomy Obervatory). Este catálogo representa el esfuerzo de recopilar, estandarizar y extender recursos espectroscópicos existentes puestos a disposición de la comunidad astronómica. Es la compilación de diversos catálogos de líneas espectrales, entre ellos destacan las bases de datos del JPL [25], CDMS [24] y Lovast/NIST [19]. Este catálogos almacena la información de lineas de emisión etiquetadas por astrónomos durante décadas, así Splatalogue reúne más de 5.8 millones datos de transiciones [5].

2.3. Trabajos relacionados

Los trabajos enfocados en la clasificación automática de líneas espectrales son recientes y no se extienden más allá de la última década. Por esta razón en la literatura se pueden encontrar aproximaciones en la resolución de este problema utilizando técnicas que permitan procesar grandes volúmenes de datos, es decir técnicas de Minería de Datos. Riveros (2016) [27] propone un algoritmo que utiliza un modelo *Sparse Coding* utilizando aprendizaje no supervisado; utiliza los datos de Splatalogue para entrenar los modelos, pero evalúa el poder de clasificación del algoritmo sobre datos sintéticos. K. Pichara trabaja en la clasificación de lineas espectrales desde el año 2003, trató de resolver este problema detectando picos de energía en el espectro utilizando un ajuste Gaussiano; luego realiza una comparación con los datos de Splatalogue y así evalúa el poder de clasificación. Miranda (2015) [23] propone una solución utilizando Reglas de Asociación, estudiando de esta manera las asociaciones lógicas entre líneas a lo largo de un espectro; Miranda obtiene buenos resultados, sin embargo utiliza un conjunto de datos relativamente pequeño, de aproximadamente 2 millones

de líneas espectrales. A. Barrientos realiza su examen de doctorado sobre aplicaciones de Máquinas de Aprendizaje para clasificar lineas espectrales utilizando Máquinas de Soporte Vectorial y Redes Neuronales Artificiales para clasificar transiciones en espectros sintéticos; Barrientos logra obtener muy buenos resultados para ANN y SVM, obteniendo este último algoritmo una menor precisión. Finalmente uno de los trabajos más recientes, y sobre el cual se respalda esta propuesta corresponde al método propuesto por M. Mendoza (2016) [22] en donde se utilizan herramientas de Minería de Texto para resolver el problema. Mendoza, al igual que Barrientos, plantea una forma de modelar los espectros utilizando solo dos características: frecuencia y energía (intensidad). Para realizar la clasificación utiliza Splatalogue como un conjunto de datos de entrenamiento para ajustar modelos de mezclas o *topic models* capaces de capturar la coocurrencia de transiciones en ventanas espectrales provenientes de observaciones con gran nivel de resolución.

2.4. Modelos de Mezclas

Los Modelos de Mezclas o Topic Models son una aplicación del mundo de la Minería de Texto (*Text Mining*, TM) y las Máquinas de Aprendizaje (*Machine Learning*, ML). Se presentan como modelos generativos probabilísticos capaces de generar representaciones de dimensionalidad reducida; se ajustan parámetros y variables latentes sobre conjuntos de datos de entrenamiento.

Un proceso generativo es un algoritmo que describe cómo se obtuvo un resultado. Por ejemplo, uno podría describir el proceso generativo de lanzar un dado: uno de los lados es seleccionado de una distribución multinomial con probabilidad $1/6$ para cada una de las caras. Estos procesos de naturaleza aleatoria son de gran utilidad para representar los procesos de generación de texto [13].

Al proveer mecanismos para presentar los resultados en términos probabilísticos, estos modelos pueden ser analizados por herramientas estadísticas estándar. Más aún, métodos Bayesianos pueden ser utilizados para realizar estimación de parámetros.

Una de las aplicaciones corresponde a capturar la coocurrencia de términos en un conjunto de

documentos y de esta forma determinar patrones sobre datos de documentos [13]; situación análoga a la de encontrar una transición molecular, definida por una frecuencia e intensidad de emisión, en un espectro de alguna región de interés (ROI).

2.4.1. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) [17] es una herramienta que utiliza un modelo generativo probabilístico para realizar separación de mezclas, por esta razón se pueden utilizar herramientas estadística estándar para realizar la estimación de parámetro. La idea central se basa en modelar conjuntos de texto junto a una variable latente. De esta forma se logra reducir la dimensionalidad en la representación de documentos de un Corpus, pasando de una representación de matriz término-frecuencia a una distribución de tópicos por documento y términos asociados a estos tópicos. Esta herramienta posee muchas aplicaciones, principalmente usada en recuperación de información, *natural language processing* (NLP), minería de texto y otras áreas relacionadas.

pLSA postula que un documento d y una palabra w_n son condicionalmente independientes dado una variable latente según la ecuación 2.6:

$$P(w_j, d_i) = p(d_i) \sum_{k=1}^{K} p(w_j \mid z_k) p(z_k \mid d_i) \tag{2.6}$$

En otras palabras, pLSA un modelo de variables latentes que modela la coocurrencia de los datos en función de estas variables. Cada variable no observada z_k se asocia con cada observación w_j.

A pesar de que pLSA se presenta como una buena base para realizar separación de mezclas, presenta dos problemas. Primero contiene una gran cantidad de parámetros que crece linealmente con el número de documentos por lo que se tiende a realizar sobreajuste en los datos de entrenamiento. Segundo, no existe una forma natural de computar la probabilidad de un documento que no pertenezca a los datos de entrenamiento [13].

2.4.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [10] se presenta como una mejora de pLSA. Blei establece un supuesto distribucional sobre la mezcla de tópicos por documento y otra sobre la distribución de términos por tópico. De esta forma se logra reducir el número de parámetros a ser aprendidos y además se provee un método claro para calcular la probabilidad de documentos no observados, es decir realizar inferencia sobre nuevos datos. En este modelo cada ítem de una colección es modelado como una mezcla finita sobre un conjunto de variables latentes, condicionadas a supuestos de distribución.

LDA posee gran cantidad de aplicaciones y es de gran utilidad para tareas como clasificación de documentos, detección de novedades, cálculos de similitud, entre otras. Suponga que se entrena un modelo LDA con textos de revistas científicas que hablan de química, biología y física; y se obtiene la siguiente distribución de palabras por tópico:

Tópico 1: molécula - 40 %, reacción - 20 %, temperatura - 10 %, ...

Tópico 2: célula - 30 %, adn - 25 % temperatura - 5 %, ...

Tópico 3: fuerza - 20 %, velocidad - 20 %, temperatura - 15 %, ...

Suponga ahora que se quiere encontrar la distribución de tópicos de un nuevo artículo d. Aplicando el modelo descrito al nuevo artículo se puede generar una distribución similar a la siguiente:

Texto	$P(z_1 \mid d)$	$P(z_2 \mid d)$	$P(z_3 \mid d)$
Las moléculas interestelares deben ser analizadas en función de la velocidad y temperatura de la fuente observada	50 %	10 %	40 %

Donde cada $P(z_i \mid d)$ corresponde a la probabilidad de que el documento d sea generado por el tópico z_i.

Modelo Generativo

El proceso generativo que LDA realiza para modelar cada documento de una colección es el siguiente:

1. Escoger cantidad de palabras $N \sim Poisson(\xi)$
2. Escoger distribución de tópicos $\theta \sim Dir(\alpha)$
3. Por cada una de las palabra $w_n, n \in \{1, ..., N\}$:
   a) Escoger un tópico $z_n \sim Multinomial(\theta)$
   b) Escoger distribución de palabras $\phi \sim Dir(\beta, z_n)$
   c) Escoger una palabra $w_n \sim Multinomial(\phi)$

Como se puede apreciar en la figura 2.3, LDA puede ser representado como un modelo de 3 niveles, dónde α y β son parámetros de la colección o hiper-parámetros, la variable θ a nivel de documentos y finalmente las variables z y w a nivel de palabras. Los hiper-parámetros α y β controlan la distribución de tópicos por documento y la distribución de palabras por tópico respectivamente. Bajo este modelo los documentos pueden ser asociados a múltiples tópicos, factor clave para realizar identificar múltiples líneas espectrales en espectros astronómicos.

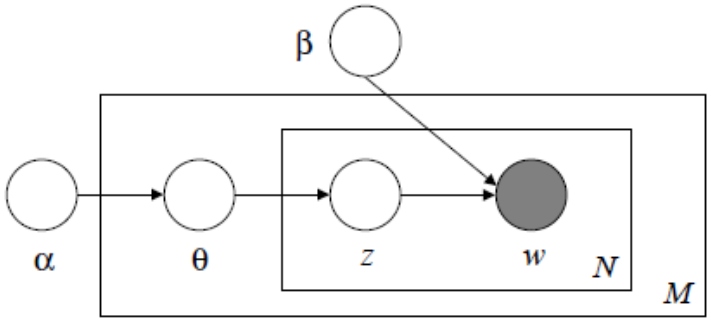


Figura 2.3: Representación gráfica del modelo generativo de LDA. El conjunto exterior M corresponde al conjunto de documentos de la colección de datos, el rectángulo interior representa el conjunto de instancias de cada palabra seleccionada por un tópico.

Dados los parámetros α y β, la distribución conjunta de una mezcla de tópicos θ, un conjunto de K tópicos z y un documento d con N palabras se define como:

$$p(\theta, z, d \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{2.7}$$

La probabilidad $p(\theta \mid \alpha)$ corresponde a la densidad de probabilidad de una variable aleatoria k-dimensional, generada desde una distribución de Dirichlet. Así θ representa la distribución de tópicos por documento y puede tomar valores en el *simplex* de dimensiones $k - 1$. Su densidad de probabilidad es la siguiente:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_i - 1} \cdots \theta_k^{\alpha_k - 1} \tag{2.8}$$

Estimación de parámetros

Para realizar el entrenamiento del modelo y estimar los parámetro se utilizan herramientas estadísticas para realizar inferencia Bayesiana. Dado una colección de M documentos, se desea encontrar los parámetros α y β que maximizan la verosimilitud de los datos:

$$\ell(\alpha, \beta) = \sum_{i=1}^{M} \log p(d_i \mid \alpha, \beta) \tag{2.9}$$

La distribución marginal $p(d \mid \alpha, \beta)$ puede ser derivada de la ecuación 2.7. Integrando sobre θ y sumando sobre z se obtiene:

$$p(d \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \tag{2.10}$$

La ecuación 2.10 puede ser escrita en términos de los parámetros del modelo. V corresponde al tamaño del espacio de características (palabras distintas) y w_j^d corresponde a la cantidad

de veces que la palabra w_j aparece en el documento d:

$$p(d \mid \alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \left(\prod_{j=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_j^d} \right) d\theta \tag{2.11}$$

Luego, para poder realizar inferencia con este modelo es necesario calcular la distribución a posterior de las variables latentes θ y z para un documento d:

$$p(\theta, z \mid d, \alpha, \beta) = \frac{p(\theta, z, d \mid \alpha, \beta)}{p(d \mid \alpha, \beta)} \tag{2.12}$$

La ecuación anterior es imposible de calcular debido al acoplamiento entre θ y β. El espacio de búsqueda se realiza sobre una productoria en la cual habría que optimizar simultáneamente en i y en j. Por esto Blei (2003) ofrece una propuesta de inferencia variacional que busca maximizar una cota inferior de la verosimilitud de los datos, considerando una familia de distribuciones indexadas por un conjunto de parámetros variacionales γ y ϕ. De esta forma se rompe el acoplamiento entre θ y β. La familia de distribuciones incorporadas es la siguiente:

$$q(\theta, z \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n) \tag{2.13}$$

El problema de encontrar una cota inferior que maximice la verosimilitud es equivalente a minimizar la divergencia de Kullback-Leibler (KL), una medida respecto a la similitud de dos distribuciones P y Q:

$$\underset{\gamma, \phi}{\operatorname{argmin}} \, D(q(\theta, z \mid \gamma, \phi) \parallel p(\theta, z \mid d, \alpha, \beta)) \tag{2.14}$$

La optimización anterior puede ser calculada mediante iteración de punto fijo de los parámetros variacionales:

$$\phi_{ni} \propto \beta_{iw_n} e^{E_q[log(\theta_i) \mid \gamma]} \tag{2.15}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{2.16}$$

$$E_q[log(\theta_i) \mid \gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \tag{2.17}$$

Utilizando el algoritmo Expectation Maximization (EM), se maximiza una cota inferior de la verosimilitud de los datos, descrita en la ecuación 2.9, con respecto a los parámetros variacionales, luego para valores fijos de estos parámetros se maximizan los valores de α y β. Los pasos del algoritmo se describen a continuación:

1. **Expectation Step.** Para cada documento se optimizan los valores de los parámetros variacionales γ_d y ϕ_d, que maximizan la verosimilitud de los datos.
2. **Maximization Step.** Se actualizan los parámetros del modelo α y β en función de los parámetros variacionales que maximizan las estimaciones de máxima verosimilitud de los datos.

Los pasos se repiten hasta converger a una cota inferior de la máxima verosimilitud de los datos descrita por la ecuación 2.9.

Se aprecia que LDA realiza un aprendizaje no supervisado debido a que modela el conjunto de documentos sin etiquetas asociadas. El algoritmo reduce la dimensionalidad de los datos y luego realiza agrupación suavizado o *soft clustering* sobre la nueva representación, de tal manera que cada documento puede estar asociado en múltiples grupos. Esto es fundamental para la identificación de múltiples líneas espectrales presentes en un documento. [13].

2.4.3. Supervised Latent Dirichlet Allocation

Supervised Labeled Latent Dirichlet Allocation (sLDA) es un modelo de mezclas que incorpora supervisión al modelo generativo. En otros modelos de tópicos solo las palabras del

documento son modeladas como datos observados. Para poder agregar supervisión al modelo de mezclas, D. Blei y J. McAuliffe (2008) [21] proponen sLDA como una extensión a LDA, una variación que puede inferir los tópicos implícitos en una colección de datos, pero basándose en que cada documento es vinculado con una respuesta o etiqueta. Así se puede inferir directamente la estructura latente de nuevos documentos utilizados en inferencia.

sLDA postula que cada etiqueta es generada de la distribución de mezclas empírica de un documento. En este modelo se modelan los documentos y etiquetas de manera conjunta, de manera de encontrar tópicos latentes que permitirán mejorar la predicción de etiquetas para documentos futuros no etiquetados. Bajo este modelo la distribución marginal de un documento y su etiqueta queda definida por:

$$p(d, y \mid \alpha, \beta, \eta, \sigma^2) = \int p(\theta, \alpha) \sum_{z_n} \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) p(y \mid z_n, \eta, \sigma^2) d\theta \qquad (2.18)$$

2.4.4. Labeled LDA

Labeled Latent Dirichlet Allocation (L-LDA) es una extensión al modelo LDA que puede describir el proceso generativo de una colección de documentos con múltiples etiquetas. D. Ramage (2009) [26], realiza esta extensión definiendo una correspondencia entre los tópicos latentes y las etiquetas de los documentos. De esta forma L-LDA es capaz de aprender las correspondencia entre etiquetas y palabras.

Este modelo es de gran utilidad en recuperación de información y análisis de colecciones de datos. Por ejemplo, un usuario que busca documentos con una etiqueta en particular quiere un resultado con los documentos que estén más relacionados con la etiqueta; o un astrónomo que busca aquellas observaciones que tengan mayor probabilidad de contener transiciones de algún elemento químico en particular.

L-LDA incorpora supervisión restringiendo al modelo para solo aprender los tópicos correspondientes a un conjunto de etiquetas Λ para cada documento. En la figura 2.4 se puede observar la relación entre el conjunto Λ y la distribución de tópicos θ. En este modelo cada documento d es representado como un tupla que contiene una lista de palabras $\mathbf{w} = (w_1, \ldots, w_{N_d})$ y una lista binaria de presencia de tópicos $\Lambda = (l_1, \ldots, l_L)$, con cada $w_i \in \{1, \ldots, V\}$ y cada $l_k \in \{0, 1\}$.

El proceso generativo es idéntico al de LDA, a excepción del paso 2, en el que se obtiene la distribución de tópicos por documento θ. L-LDA restringe θ a ser definido solo por los tópicos presentes en el conjunto de etiquetas Λ del documento:

1. Escoger cantidad de palabras $N_d \sim Poisson(\xi)$
2. Para cada tópico:
   a) Escoger $\Lambda_k \sim Bernoulli(\Phi_k)$
3. Generar $\alpha' = L \times \alpha$ y
4. Escoger distribución de tópicos $\theta \sim Dir(\alpha')$
5. Por cada una de las palabra $w_n, n \in \{1, ..., N_d\}$:
   a) Escoger un tópico $z_n \sim Multinomial(\theta)$
   b) Escoger distribución de palabras $\phi \sim Dir(\beta, z_n)$
   c) Escoger una palabra $w_n \sim Multinomial(\phi)$

Se define un vector de etiquetas de documento $\lambda = \{k \mid \Lambda_k = 1\}$, esto permite definir además un matriz de proyección de etiquetas L de tamaño $M \times K$, con $M = |\lambda|$:

$$L_{ij} = \begin{cases} 1 & \text{si} \quad \lambda_i = j \\ 0 & \text{e.t.o.c.} \end{cases} \tag{2.19}$$

Esta matriz de proyección se utiliza para obtener el parámetro α' de menor dimensionalidad que α. Así θ es obtenido desde una distribución de Dirichlet con parámetro α':

$$\alpha' = L \times \alpha = (\alpha_{\lambda_1}, \ldots, \alpha_{\lambda_M}) \tag{2.20}$$

Con el nuevo parámetro α se restringe a la distribución de Dirichlet a los tópicos pertenecientes a las etiquetas del documento λ

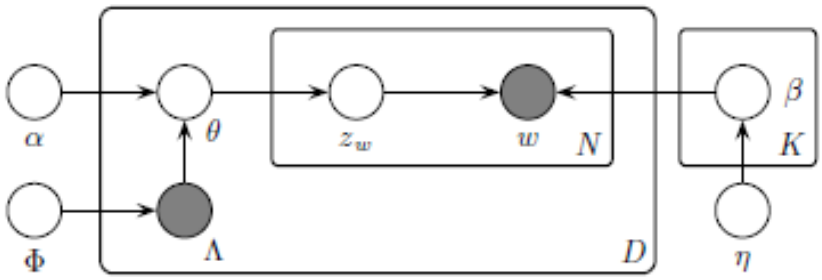


Figura 2.4: Representación gráfica del modelo generativo de L-LDA. Para el modelo propuesto D representa el conjunto de espectros, N el conjunto de emisiones del espectro o transiciones y K el conjunto de especies moleculares. Se puede interpretar como cada transición es generada por una especie molecular.

Estimación de parámetros

A diferencia de la estimación de parámetros realizada con el algoritmo EM, que intenta estimar los valores de θ y ϕ; para el algoritmo L-LDA propuesto por D. Ramage (2009) se utiliza Collapsed Gibbs Sampling un procedimiento basado en variables condicionales de los estados anteriores [16] y se desea estimar la probabilidad a posterior de la asignación de tópicos a palabras:

$$P(z \mid w) = \frac{P(w, z)}{\sum_k P(w, z)} \tag{2.21}$$

Para computar $P(z \mid w)$ se necesita evaluar una distribución de probabilidades sobre un espacio muy grande de variables discretas. Para resolver esto se utilizan cadenas de Markov Monte Carlo, cadenas que simulan procesos de memoria larga. Ahora se busca estimar la probabilidad $P(z \mid w)$ en función de las asignaciones actuales realizando iteraciones hasta converger a las probabilidades que maximicen la verosimilitud de los datos:

$$P(z_i = j \mid \boldsymbol{z_{-i}}, w_i, d_i) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \times \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{2.22}$$

Donde C^{WT} corresponde a una matriz palabra - tópico de tamaño $V \times K$ y $C^{WT}_{w_i j}$ corresponde al elemento $c_i j$ de la matriz, que indica la cantidad de veces que la palabra i aparecen asignada al tópico j. El término $C^{DT}_{d_i j}$ corresponde a la matriz documento - tópico de tamaño $M_d \times K$, con M_d la cantidad de documentos de la colección; cada elemento de esta matriz representa la cantidad de veces que el tópico j ocurre en el documento d_j.

Por cada iteración o estado de una cadena de Markov, es decir, por cada asignación de tópicos a un conjunto de palabras, se pueden estimar los parámetros θ y ϕ:

$$\theta_{d_i j} = \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{2.23}$$

$$\phi_{w_i j} = \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \tag{2.24}$$

Una vez que el modelo aprende las distribuciones multinomiales de tópicos β, se puede realizar inferencia sobre nuevos documentos utilizando Gibbs Sampling, restringido al conjunto de etiquetas establecido en entrenamiento, para determinar la asignación etiquetas a palabras z.

Capítulo 3

Metodología

Para construir el algoritmo de clasificación de líneas espectrales propuesto se deben realizar tres tareas: (i) obtener datos de transiciones espectrales, (ii) entrenar modelos L-LDA, (iii) procesar espectros generados desde cubos de datos o *surveys* de líneas espectrales y (iv) realizar inferencia con los modelos L-LDA sobre los datos procesados y evaluar el poder de predicción.

El código de este proyecto se presentará de forma *open-source*, disponible en un repositorio Github [1] para que la comunidad pueda replicar y mejorar el algoritmo de clasificación. Por esta razón las datos y herramientas utilizadas también deben ser de naturaleza *open-source*.

3.1. Datos

3.1.1. Splatalogue

Uno de los principales objetivos de Splatalogue es disponibilizar los datos de manera intuitiva tanto para usuarios novatos y expertos. Así, Splatalogue permite realizar consultas utilizando diversos filtros para obtener transiciones clasificadas por tipo de objeto. Estos

[1] `https://github.com/sebastian-aranda/Labeled-LDA-Spectral-Line-Classification`

objetos se dividen en: *Planetary Atmosphere*, *Active Galactic Nucleus/Proto Planetary Nebula/Planetary Nebula* (AGB/PPN/PN), *Hot Cores*, *Diffuse Clouds*, *Dark Clouds*, *Comets* y *Extragalactic*. Además Splatalogue permite realizar búsquedas por bandas de frecuencia por lo que se pueden obtener transiciones registradas para distintas bandas de los receptores de ALMA. Información de las características de estas colecciones se detalla en la tabla 3.1.

Cuadro 3.1: Resumen de las características de cada modelo por tipo de objeto y banda receptora de ALMA (AB). Para cada modelo se detalla la cantidad de especies moleculares y la cantidad de transiciones totales. Se omiten los datos de las AB 3,4,5,8,9 y 10.

Modelo	Especies	Transiciones
Planetary	20	25778
Hot Cores	85	757475
AGB/PPN/PN	71	245473
Dark Clouds	58	306246
Diffuse Clouds	24	150400
Extragalactic	36	74542
Comets	30	115404
AB 6 (211-275 GHz)	133	98086
AB 7 (275-373 GHz)	132	131698

Para cada uno de los tipos de objetos estelares y las distintas bandas de frecuencia de ALMA, se obtiene el listado de transiciones. Para cada transición se almacenan sus datos: nombre de la especie, frecuencia de reposo f [GHz], estado de energía inferior (E_l) [K] y el nombre del catálogo donde fueron registradas (JPL, CDMS, etc).

Splatalogue registra las frecuencias de emisión medidas en GHz con hasta 5 dígitos de precisión. Esto permite aprovechar el nivel de detalle del catálogo para identificar líneas espectrales en observaciones con gran nivel de resolución, como los cubos de datos de ALMA. Esta característica permite evaluar el algoritmo sobre distintos esquemas de resolución o *channeling*.

3.1.2. Surveys

La gran tecnología que poseen los nuevos observatorios ha permitido incrementar el nivel de resolución de los datos espectrales obtenidos. Esto ha motivado a científicos a realizar inspecciones en zonas del cielo, en bandas de frecuencia extensas y con abundancia de líneas espectrales; P. Schilke (2001) [28] y C. Comito (2005) [12] han realizado este tipo de investigación con el objetivo de identificar y medir la abundancia de moléculas, en una zona del espacio que se caracteriza por la formación de estrellas, Orion-KL. Este tipo de regiones presentan reacciones químicas complejas que dan origen diversas especies moleculares y a sobre posición parcial de líneas espectrales.

Como primera aproximación se evalúa el poder predictivo del algoritmo realizando clasificación multi- etiqueta. Para esto se recopila el conjunto completo de transiciones registradas por P. Schilke et. al. [28] en la zona de formación de estrellas Orion-KL. Se evalúan distintas regiones de frecuencias del *survey* para así medir el poder de predicción sobre ventanas espectrales con cantidades fijas de transiciones.

Se desea evaluar el algoritmo analizando distintos rangos de frecuencias, para así analizar el poder de predicción con documentos de distinto tamaño, saturados con transiciones. Sin embargo, realizar esta clasificación no es trivial debido a la complejidad de la tarea, la mayor dificultad recae en la identificación de moléculas complejas como Methanol, que poseen cientos de transiciones en el espectro milimétrico/sub-milimétrico, estas moléculas influyen en la clasificación de líneas más débiles [20]. Por esta razón se debe tener mucho cuidado en la identificación de moléculas en *surveys* de líneas espectrales.

3.1.3. FITS

Los archivos FITS son el formato de archivos estándar más utilizado en la disciplina de la astronomía. A pesar de que existen ligeras diferencias en la manera en que cada observatorio almacena sus datos, los archivos FITS definen una estructura general para guardar los datos de una observación. Así un archivo FITS está compuesto de una cabecera que almacena los metadatos de una observación y una sección de datos binarios que representan la imagen

capturada. La figura 3.1 muestra el contenido de una imagen FITS.

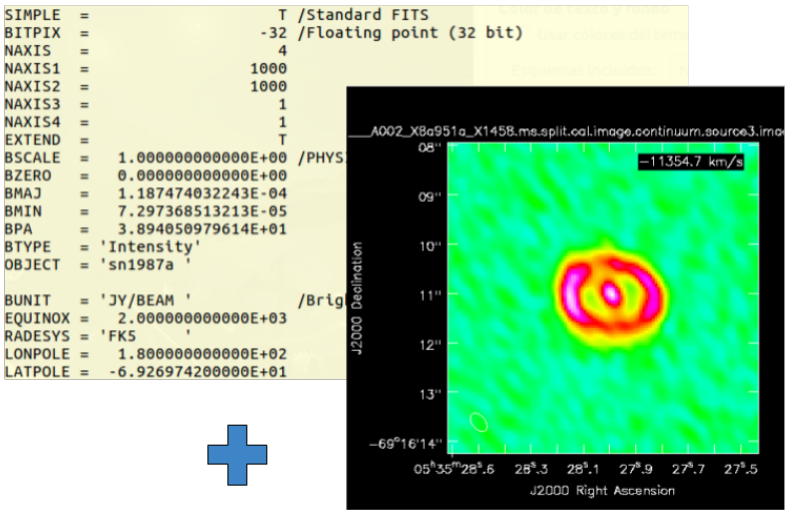


Figura 3.1: Ejemplo del contenido de un archivo FITS. Se presenta el *header* o metadatos del archivo a un lado izquierdo y el contenido de la imagen a un lado derecho. Proyecto ALMA #2013.1.00280.S.

Producto del desarrollo de tecnologías de vanguardia, en radioastronomía los archivos FITS almacenan cubos de datos, imágenes con un eje espectral (frecuencia, longitud de onda o velocidad) adicional. El objetivo central de la propuesta es presentar un método novedoso en la identificación de líneas espectrales para cubos de datos de ALMA.

La radioastronomía se enfoca en la porción del espectro electromagnético correspondiente a las ondas cuyas longitudes de onda pueden variar entre 0.001 - 30 metros aproximadamente. Debido a que estos instrumentos observan en un amplio rango de longitudes de onda, las observaciones realizadas generan productos científicos de gran tamaño conocidos como cubos de datos. Estos cubos representan una imagen con dos ejes espaciales y tercer eje espectral [14].

Debido al gran volumen de datos que representan estas observaciones es necesario desarrollar herramientas de software capaces de realizar análisis automático y eficiente. Trabajos relacionados con la indexación de objetos estelares en cubos de datos es realizado por M. Araya et. al. [6]. Aquí se proponen herramientas que permiten agilizar la identificación de objetos sobre grandes volúmenes de datos.

3.2. Algoritmo

En esta propuesta se busca modelar los cubos de datos como conjuntos de palabras, o documentos; cada documento representa de esta forma un espectro de un punto o una región de interés de una observación. Luego se entrenan distintos modelo de mezclas con el algoritmo L-LDA, utilizando datos de Splatalogue. Finalmente se procede a realizar inferencia con los modelos entrenados sobre la representación generada.

3.2.1. Entrenamiento de Modelos

Se descargan todas las transiciones registradas en la base de datos de Splatalogue. Con estos datos se obtienen 9 colecciones de transiciones distintas, 7 por cada tipo de objeto y 2 por las bandas de ALMA 6 y 7 (ver tabla 3.1).

Se obtienen ejemplos de entrenamiento para cada uno de los filtros: por cada transición se registra la frecuencia [GHz] y energía [K] de emisión. Haciendo el análogo con análisis de texto, cada frecuencia representaría una palabra, o una característica del espacio de características; la energía de emisión correspondería al peso o importancia de la característica.

Se introduce el parámetro *channeling* para controlar el nivel de resolución de los espacios de características generados. Este parámetro controla el nivel de discretización de las frecuencias de emisión; se definen 5 niveles de resolución x, que consideran el valor de las

frecuencias f medidas en [GHz] con hasta 5 dígitos decimales. La función *channeling* corresponde a:

$$channeling(f) = \lfloor f \cdot 10^{x} \rfloor \tag{3.1}$$

De esta forma cada característica es representada como un mapeo de la función *channeling*. El peso de cada característica es calculado utilizando una función que imita la función de importancia término-frecuencia tf utilizada en análisis de texto, en particular se calcula la cantidad de veces que se repite cada frecuencia de transición en función de su intensidad de emisión o energía E. Esta relación corresponde a:

$$tf(f) = \lceil \log_2(E + 1) \rceil \tag{3.2}$$

Notar que el operador logaritmo de la ecuación 3.2 se incorpora para atenuar la intensidad de transiciones muy energéticas. Sin el operador estaríamos saturando los ejemplos de entrenamiento con estas transiciones y el modelo sería más sensible a experimentar sobreajuste (overfiting).

Con las definiciones anteriores es factible generar una representación válida para entrenar modelos L-LDA. Los documentos de entrenamiento son generados utilizando los operadores tf y *channeling*. Se genera una colección de documentos por cada uno de los filtros de Splatalogue. Cada documento de la colección representa el conjunto de transiciones de cierta especie molecular, modeladas por los operadores descritos. Cada documento es construidos de la siguiente manera:

1. Para una especie molecular se obtiene el listado completo de transiciones en forma de tuplas $\langle frecuencia[GHz], energía[K] \rangle$.

2. Para cada transición se aplica el operador *channeling* al valor de frecuencia para controlar el nivel de resolución del espacio de características.

3. Cada frecuencia mapeada con el operador *channeling* representa una palabra (w_n) que

es insertada en el documento la cantidad de veces indicadas por la función de peso tf. La cantidad de repeticiones es proporcional a la intensidad de la transición.

4. Finalmente se tiene un listado de tuplas de la forma $\langle channeling(frecuencia), tf(energía)\rangle$ que representa un documento con las transiciones de una especie molecular. correspondiente.

Para realizar el entrenamiento de los modelos, se utiliza una implementación de L-LDA en Java llamada JGibbLabeledLDA [2]. Esta librería realiza la estimación de parámetros utilizando Gibbs Sampling. Los hiper-parámetros son calculados utilizando un estándar que ha tenido muy buenos resultados con diversas colecciones de datos [18]:

$$\alpha = \frac{50}{k} \tag{3.3}$$

$$\beta = \frac{K}{M} \tag{3.4}$$

Cada modelo se entrena con los hiper parámetros calculados en función de la cantidad de especies moleculares K (cantidad de tópicos) y la cantidad transiciones M (tamaño del espacio de características). La cantidad de especies K las define el filtro o banda de alma establecida y la cantidad de transiciones M las define el nivel de resolución o *channeling* escogido sobre el mismo filtro.

3.2.2. Expansión de términos

Los modelos son entrenados utilizando datos de Splatalogue (i.e. solo información de transiciones), sin embargo, se desea evaluar la capacidad de predicción de los modelos con espectros de cubos de datos, por esto es necesario realizar cierto ajuste a los datos de entrenamiento para mejorar la capacidad de generalización del algoritmo.

[2] https://github.com/myleott/JGibbLabeledLDA

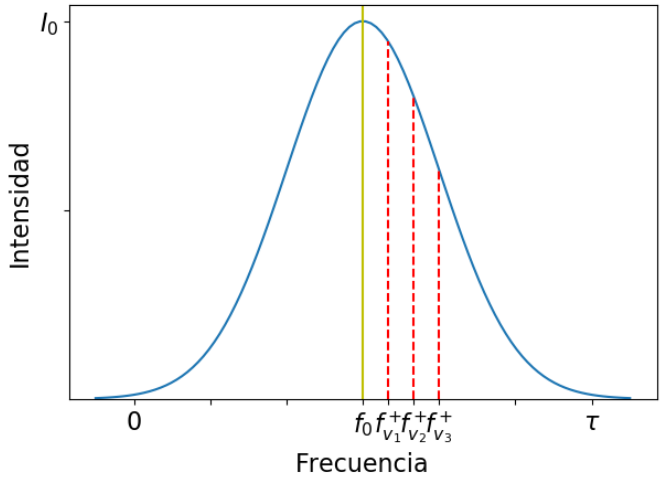


Figura 3.2: Ejemplo de la expansión de vecindad para una transición con frecuencia f_0 e intensidad I_0; se establece la cantidad de vecinos en $\chi = 3$. Por simplicidad se muestra solo la mitad derecha de la expansión.

Las líneas espectrales marcan las frecuencias puntuales en las que se emiten ondas electromagnéticas, sin embargo al observar objetos en astronomía se puede apreciar que líneas de emisión o absorción son representadas como picos o valles de intensidad con un ancho particular. Una de las razones físicas de este fenómeno es producido por el ensanchamiento de Doppler, efecto causado por las variaciones de velocidad y estados de excitación en las partículas de un objeto que emite radiación; temperaturas más altas implican un mayor ensanchamiento de las líneas espectrales. Este fenómeno se aprecia en la figura 2.2

La propuesta consiste en realizar una expansión del vocabulario de cada colección de datos, agregando palabras que se encuentren en la vecindad de cada transición. Esta vecindad se establece modelando una campana de Gauss centrada en cada una de las transiciones. Se define el parámetro τ que modela el soporte de la expansión, en otras palabras, el ancho base de una transición. Se asume que los perfiles generados por el ensanchamiento de Doppler se comportan como una distribución normal cuya desviación estándar corresponde a:

$$\sigma = \frac{\tau}{6} \tag{3.5}$$

De esta manera la expansión se realiza agregando $2 \times \chi$ transiciones vecinas a cada línea espectral. El parámetro χ define cuantos vecinos se agregan a cada lado de la transición puntual. En la figura 3.2 se puede apreciar como se modelan los vecinos de una transición con frecuencia f_0 e intensidad I_0.

La expansión de términos define una nueva colección de transiciones, con más cantidad de información en función del nivel de soporte τ y la cantidad de vecinos a agregar χ. De esta forma se pueden entrenar modelos adicionales para cada uno de los filtros de Splatalogue.

3.2.3. Procesamiento de cubos de datos o FITS

Los cubos de datos generados por ALMA son almacenados en archivos formato FITS [3]. Para poder leer y procesar estos archivos se utilizará una paquete de python utilizado por astrónomos conocido como Astropy [4]. Esta librería permite leer los metadatos y datos desde archivos FITS. Posee una extensa documentación con ejemplos de procesamiento de datos; además existe una comunidad astro-informática que se preocupa de mejorar la librería constantemente.

Utilizando Astropy se obtienen los siguientes metadatos desde la cabecera del cubo de datos:

- NAXIS1, NAXIS2, NAXIS3: cantidad de componentes de cada eje espacial 1, eje espacial 2 y eje espectral, respectivamente.
- RESTFRQ: dependiendo del tipo de observación, representa la frecuencia de una línea espectral en interés o el centro de la ventana espectral. Este valor está medido en [Hz].
- CRVAL3 valor del primer elemento del eje espectral, medido en unidades de frecuencia [Hz], velocidad [km/s] o longitud de onda [m].
- CRDELT3: distancia entre cada elemento del eje espectral, medido en la misma unidad que CRVAL3

[3] http://www.alma.inaf.it/images/ArchiveKeyworkds.pdf
[4] http://www.astropy.org/

- BMAJ, BMIN: rayo de recuperación mayor y menor respectivamente.

Luego se obtienen las regiones de interés (ROIs) del cubo de datos utilizando la clase DBSCAN (Density-Based Spatial Clustering of Applications with Noise) de la librería de python scikit-learn [5]. DBSCAN permite encontrar clusters de distinta forma en datos con regiones de densidad similar y para datos con ruido presente [15]. La librería Scikit-Learn también es de naturaleza *open-source* y cuenta con un amplia documentación, detalles del algoritmo DBSCAN se pueden encontrar en la documentación de la librería.

Para utilizar el algoritmo en un cubo de datos se realizan los siguientes pasos:

1. Se obtiene el momento 0 [6] del cubo.
2. Se identifica la región de interés (*cluster*) cuya suma de píxeles sea la más alta. Ver figura 3.3
3. Para cada canal del cubo de datos se obtiene la suma de las intensidades de los píxeles pertenecientes a la región de interés seleccionada.

Con lo anterior se genera un espectro de intensidades, pero es necesario realizar correcciones de las unidades de intensidad y de las unidades de frecuencia. Primero se deben transformar las unidades de intensidad utilizadas en radio astronomía, los cubos de datos de ALMA se miden con la unidad de densidad de flujo $Jy/beam$, pero la intensidad de emisión de las lineas espectrales catalogadas en Splatalogue se miden en temperatura de brillo T_b. Para corregir esta diferencia de escalas, se utiliza la ley de Rayleigh-Jeans según la ecuación 3.6:

$$T_b = 1{,}36\frac{\lambda^2}{\theta^2}S \tag{3.6}$$

Donde la longitud de onda λ es medida en cm, el ancho del rayo de potencia media θ o Half Power Beam Width (HPBW) se mide en segundos de arco $arcsec$ y la densidad de flujo S en $mJy/beam$. La longitud de onda λ se obtiene incorporando la frecuencia observada en la

[5] http://scikit-learn.org/stable/

[6] El momento 0 de un cubo de datos corresponde a la suma de todos los píxeles a lo largo del eje espectral

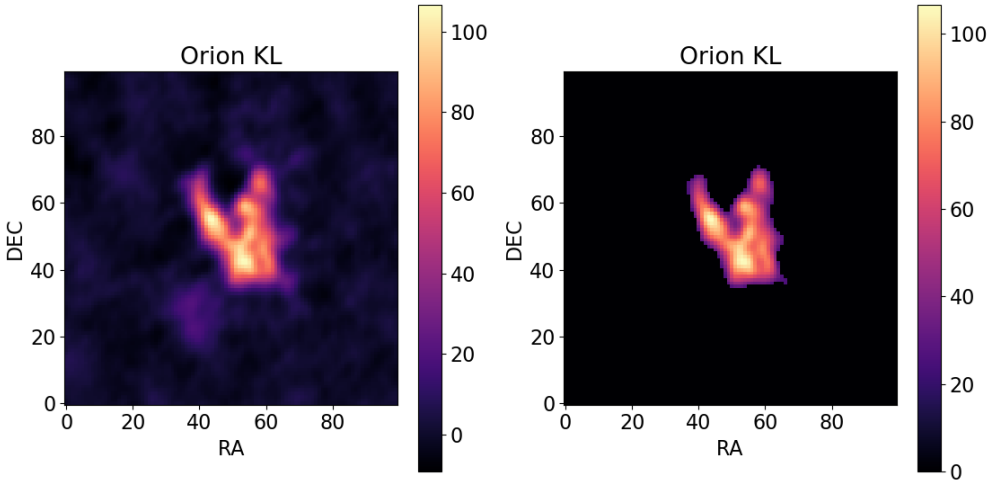


Figura 3.3: Detección de regiones de interés para Orion KL. A un lado izquierda se presenta el momento 0 del cubo de datos y a un lado derecho se presenta la región de interés seleccionada por el algoritmo DBSCAN. Proyecto ALMA #2011.0.00009.SV.

relación 2.1, el HPBW se calcula multiplicando los rayos de recuperación mayor y menor (BMAJ y BMIN), la densidad de flujo S corresponde a la intensidad medida en la región de interés obtenida del cubo de datos.

Una vez realizada la corrección de intensidades se realiza la corrección de frecuencias generadas por el efecto Doppler. Para poder poder calcular la frecuencia de reposo del espectro se utiliza la ecuación 2.3. Se utiliza el valor de la frecuencia de reposo f_0 y la frecuencia observada f_{obs} del canal central, estos valores se puede incluir en la ecuación 2.4 para obtener el valor del redshift del objeto:

$$z = \frac{f_0 - f_{obs}}{f_{obs}}$$

Luego para cada canal del espectro se realiza la corrección de frecuencias con la siguiente

función:

$$shift(f) = f(1 + z) \tag{3.7}$$

Posteriormente se debe aplicar el operador *channeling* descrito en la sección de entrenamiento, en la ecuación 3.1. Este operador se aplica para controlar el nivel de resolución del espacio de características de cada modelo. También se debe realizar un mapeo de frecuencias aplicando este operador a cada canal del espectro.

Por último se debe realizar el mapeo de frecuencias resultantes a características del vocabulario del modelo. Este es un paso crucial para que el algoritmo pueda funcionar correctamente, es necesario que los documentos generados posean las características del modelo con el cual se realiza inferencia. Para cumplir el requisito se construye una función de mapeo que busca la transición más parecida, es decir, cuya diferencia sea la menor:

$$fmap(f_0) = \underset{f_i \in F}{\operatorname{argmin}}(|f_0 - f_i|) \tag{3.8}$$

Donde F corresponde al conjunto de características distintas del modelo, f_0 corresponde a la frecuencia calculada luego de la corrección de Doppler y de la aplicación del operador *channeling*, f_i toma el valor de cada una de las características presentes en el modelo. El problema de esto es que debido al nivel de resolución con el que se trabajan los datos de frecuencia, se pueden presentar errores de mapeo debido a variaciones sutiles en la distancia entre las frecuencias comparadas.Para evitar este tipo de error se establece un umbral ξ que condiciona al operador descrito para realizar un mapeo de características sólo en el caso de que la distancia con la frecuencia a mapear no supere este umbral:

$$fmap'(f_0) = \begin{cases} \underset{f_i \in F}{\operatorname{argmin}}(|f_0 - f_i|) & \text{si} \quad |f_0 - f_i| < \xi \\ f_0 & \text{e.t.o.c.} \end{cases} \tag{3.9}$$

Una vez se genera el conjunto de transiciones con las frecuencias corregidas por el Efecto Doppler, truncadas por el operador *channeling* y mapeadas al vocabulario del modelo, se

procede a ingresar cada transición la cantidad de veces indicada por el operador tf descrito en la ecuación 3.2. La diferencia con el operador tf aplicado en los documentos de entrenamiento es la incorporación de un umbral que considere solo aquellas transiciones cuya energía supere este umbral.

$$tf'(E_f) = \begin{cases} \lceil \log_2(E+1) \rceil & \text{si} \quad E > \sigma \cdot \upsilon \\ 0 & \text{e.t.o.c.} \end{cases} \tag{3.10}$$

Este umbral es definido por la desviación estándar σ de las energías de las transiciones. Además se define el factor υ que regula el valor del umbral en función de la cantidad de canales del cubo de datos (NAXIS3), de esta manera se establece un umbral adaptivo, en función del tamaño del documento generado.

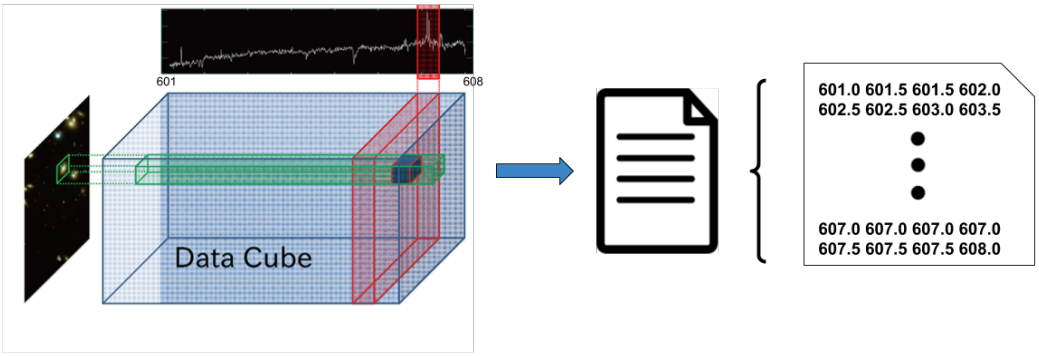


Figura 3.4: Representación gráfica de la etapa de procesamiento. Se procesa el cubo de datos generando un espectro en formato de documento desde una región de interés determinada.

3.2.4. Inferencia sobre documentos espectrales

Una vez realizados los pasos de pre-procesamiento se obtiene la representación adecuada para ser evaluada por L-LDA. Se utiliza la misma implementación utilizada para entrenar los modelos, JGibbLabeledLDA, pero ahora aplicada en la estimación de parámetros mediante inferencia posterior. El algoritmo recibe como entrada el documento generado de la etapa de procesamiento; luego se realiza inferencia sobre los nuevos datos astronomía utilizando

Gibbs Sampling, estimando la distribución de especies θ_d correspondiente:

$$\theta_d = [\theta_1^d, \theta_2^d, \ldots, \theta_k^d] \tag{3.11}$$

Cada elemento de θ_i^d corresponde a la probabilidad de que la especie molecular *i* esté presente en el espectro generado. Así se puede evaluar la presencia de una o muchas líneas espectrales en un espectro representado como documento. Para los casos en que se detecta una sola especie molecular se obtiene la probabilidad máxima de la distribución.

$$\underset{\theta_i \in \theta_d}{\text{argmax}}\{\theta_1^d, \theta_2^d, \ldots, \theta_k^d\} \tag{3.12}$$

3.2.5. Clasificación de líneas espectrales

Se desea evaluar el poder de predicción del algoritmo en dos escenarios: *surveys* de líneas espectrales y cubos de datos. Para ambos casos es necesario utilizar métricas de evaluación adecuadas, que consideren el tipo de predicción que realiza el algoritmo. Para los *surveys* de líneas espectrales es necesario evaluar la predicción sobre un conjunto de muchas especies moleculares; para la clasificación sobre cubos de datos con transiciones puntuales es necesario evaluar que algoritmo realice una predicción acertada.

Se utilizan métricas, del campo de Máquinas de Aprendizaje, conocidas como *Precision* y *Recall*. Para entender estas métricas se deben definir dos conjuntos:

- **recomendados**: son el conjunto de especies recomendadas por el modelo L-LDA para una ventana espectral, es decir, las especies a las cuales se les asigna mayor probabilidad.
- **relevantes**: son el conjunto de especies totales que pertenecen a una ventana espectral.

Al ejecutar el algoritmo para evaluar un espectro, L-LDA entrega un vector de probabilidad θ_d donde cada elemento θ_i^d indica la probabilidad de encontrar la especie *i* en el espectro

generado. Una especie se considera recomendada por el algoritmo si la probabilidad asignada supera el valor de la probabilidad más baja.

Precision

Precision es una métrica que evalúa la proporción de elementos relevantes de una recomendación. Se utiliza para determinar qué tan correcta es una recomendación.

$$Precision = \frac{|recomendados \cap relevantes|}{|recomendados|} \tag{3.13}$$

Recall

Recall es una métrica que evalúa la proporción de elementos recomendados que son relevantes. Se utiliza para determinar la completitud de una recomendación, es decir, si una recomendación posee toda la información que se espera recuperar.

$$Recall = \frac{|recomendados \cap relevantes|}{|relevantes|} \tag{3.14}$$

En el caso de los *surveys* de líneas espectrales, se generan espectros saturados con transiciones para luego evaluar el poder de predicción multi-etiqueta. Por esta razón se incorpora una variación a las métricas descritas, para evaluar el poder de predicción del algoritmo en función de la cantidad de especies consideradas en una recomendación. Se utilizan las métricas *Precision* & *Recall at k*.

Precision at k (P@k)

Precision at k mide la precisión de un listado de etiquetas sugeridas hasta la k-ésima posición. En este caso L-LLDA entrega un vector de distribución de tópicos θ, el cual es ordenado de mayor a menor probabilidad. Así se procede a obtener la precisión considerando las especies

relevantes sobre el número de especies recomendadas

$$P@k = \frac{\sum_{i=1}^{k} Rel(i)}{k} \tag{3.15}$$

Donde $Rel(i)$ corresponde a:

$$Rel(i) = \begin{cases} 1 & \text{si la especie i es relevante para el usuario} \\ 0 & \text{e.t.o.c.} \end{cases}$$

R@k se define de igual manera, es decir, se evalúa la proporción de transiciones recomendadas que son relevantes sobre el top k de recomendaciones.

Se busca que el algoritmo posea un buen nivel de precisión, es decir, que las recomendaciones realizadas sean acertadas. Además de esto se desea evaluar qué tan eficaz es el algoritmo para encontrar todas las transiciones de una ventana espectral, de tal manera de no perder información valiosa.

Capítulo 4

Experimentación

Este trabajo presenta una propuesta para clasificar líneas espectrales en datos de observaciones reales, realizando inferencia sobre estos datos utilizando como base modelos entrenados con datos obtenidos de Splatalogue. El objetivo de este trabajo es evaluar el poder de predicción del algoritmo sobre datos de transiciones recopiladas por un *survey* de líneas espectrales y sobre transiciones presentes en cubos de datos de ALMA.

La experimentación se separa en tres tareas: entrenamiento de los modelos, clasificación de transiciones en *surveys* de líneas espectrales y clasificación de transiciones en cubos de datos de radioastronomía como los generados por ALMA. En una primera etapa se busca definir y caracterizar los modelos para posteriormente evaluar su poder de predicción sobre datos reales.

Para la experimentación respectivas a los tiempos de entrenamiento y de inferencia, se utilizó un computador personal con las siguientes características:

Cuadro 4.1: Características PC utilizado para realizar la experimentación

Modelo CPU	N Núcleos	RAM
AMD A10-9600P RADEON R5, 10 COMPUTE CORES 4C+6G	10	16 GB

4.1. Modelos

Para entrenar los diferentes modelos L-LDA se utilizan los datos de Splatalogue segmentados por tipo de objeto y banda de frecuencias. Además, para cada filtro anterior se evalúan tres diferentes esquemas de resolución o *channeling* (x), que definen el nivel de resolución del espacio de características. Se utilizan los valores $x = 0$, $x = 2$, $x = 5$; un valor alto de *channeling* genera modelos con espacios de características de gran tamaño, este número de características se define como M. Para calcular el valor de los hiper-parámetros se utilizan las ecuaciones 3.3 y 3.4. Un resumen de los modelos y de sus hiper parámetros se puede ver en el Anexo 4.5, tablas 4.9 y 4.10.

Aquellos modelos con un nivel de resolución o *channeling* muy alto pueden ser aplicados a la identificación de líneas espectrales en regiones con abundancia de transiciones. Un modelo con bajo nivel de resolución no es capaz de identificar con precisión aquellas zonas con densidad de líneas, pero reducen el tiempo de entrenamiento e inferencia de los modelos; en la sección 4.1.1 se presenta un análisis de estos tiempos.

Los modelos basados en el tipo de objeto son creados para evaluar el poder de predicción en función del tipo de objeto que se evalúa. Los modelos basados en las bandas 6 y 7 de ALMA se construyen con el objetivo de generar modelos que sean capaces de predecir transiciones en rangos de frecuencia establecidos. Por esta razón, los FITS de testeo varían de acuerdo al modelo seleccionado (ver Anexo 4.4).

Se identifican dos aspectos de la construcción de los modelos: (1) es importante que las características de entrenamiento se ajusten con las características que se obtienen procesando los datos reales, por esto la aplicación del factor *channeling* debe realizarse de igual manera tanto en el entrenamiento como en la construcción de los documentos de testeo; (2) el operador término-frecuencia descrito en la ecuación 3.2 permite la evaluación de datos con intensidades de emisión bajas $< 1K$ gracias al factor de soporte que se incorpora (sumando 1 al valor de la energía).

La cantidad de transiciones de cada tópico varía significativamente en algunos casos. Por ejemplo para el filtro *Hot Cores* la especie molecular con menor cantidad de transiciones

corresponde a Isocyanide Acid con 64828 palabras (tokens), mientras que la especie Nitrosyl Hydride posee 7398317 palabras. Se espera que este sea un factor relevante al momento de clasificar líneas espectrales.

Para entrenar los modelos expandidos se realiza un ajuste en los hiper parámetros utilizados. Debido a que estos modelos poseen espacios de características mucho más grandes que los modelos estándar, se decide utilizar $\beta = 0{,}1$, el valor por defecto de este hiper parámetro; el valor de α se mantiene igual debido a que no se varía la cantidad de tópicos. Para cada uno de los filtros Splatalogue recopilados, se generan 3 modelos expandidos en función del soporte τ de la expansión. Este procedimiento se se realiza solo para los modelos con esquema *channeling* de máxima resolución ($x = 5$). A continuación se muestra el detalle de la programación del algoritmo de expansión de términos:

```
1  support = 0.01000 #Width of gaussian profile base, measured in GHz
2  neighbors = 3 #Number of neighbors to add
3
4  ldda_output_corpus = 'hot_cores_full_expanded'
5  llda_input_corpus = "hot_cores_full.dat"
6  with open(llda_input_corpus) as f:
7    mFile = open(ldda_output_corpus, 'w')
8    for i,line in enumerate(f):
9      #Getting transitions per each molecular species (each row)
10     freqs = [int(freq) for freq in line.split()[1:]]
11
12     #Setting a term-frequency dict
13     freqs_dict = defaultdict(int)
14     for freq in freqs:
15       freqs_dict[freq] += 1
16
17     #Aplying Gaussian Expansion
18     for central_freq, central_freq_count in freqs_dict.items():
19
20       #Setting Gaussian Profile
```

```
21          normal = stats.norm(central_freq,support/6)
22          for i in range(neighbors):
23            #Getting new frequency neighbors
24            new_freq_right = int(central_freq+(i+1)*(support/2)/neighbors)
25            new_freq_left = int(central_freq-(i+1)*(support/2)/neighbors)
26
27            new_count_right = int(
28                ceil(central_freq_count/normal.pdf(central_freq)
29                *normal.pdf(new_freq_right)
30                )
31            )
32            new_count_left = int(
33                ceil(central_freq_count/normal.pdf(central_freq)
34                *normal.pdf(new_freq_left)
35                )
36            )
37
38            freqs.extend([new_freq_right for i in range(new_count_right)])
39            freqs.extend([new_freq_left for i in range(new_count_left)])
40
41        #Sorting list of characteristics
42        freqs.sort()
43
44        mFile.write(line.split()[0]+" "+" ".join(map(str,freqs))+"\n")
45      mFile.close()
```

El tiempo de creación de las nuevas colecciones no es considerado en el análisis, pero se destaca la utilización de la estructura de datos *defaultdict*, provista por la librería *collections* de *python*, para crear el diccionario de término-frecuencia. La declaración de esta estructura de datos permite realizar la expansión de términos, de manera eficiente, desde el mismo archivo utilizado para entrenar los modelos L-LDA clásicos.

4.1.1. Tiempos de entrenamiento

La figura 4.1 presenta las curvas de tiempo transcurrido para el entrenamiento de los distintos modelos de channeling. Se aprecia un claro comportamiento lineal en función del tamaño de los documentos de entrenamiento; además existe un punto en donde las diferencias de tiempo se incrementan entre los distintos esquemas de *channeling*, cerca de 2500000 palabras para los modelos clásicos y 10000000 palabras para los modelos expandidos. Las diferencias de tiempo más grandes se presentan para los modelos con mayor cantidad de palabras de entrenamiento, modelos que utilizan el filtro de *Hot Cores* presentan los mayores tiempos de entrenamiento.

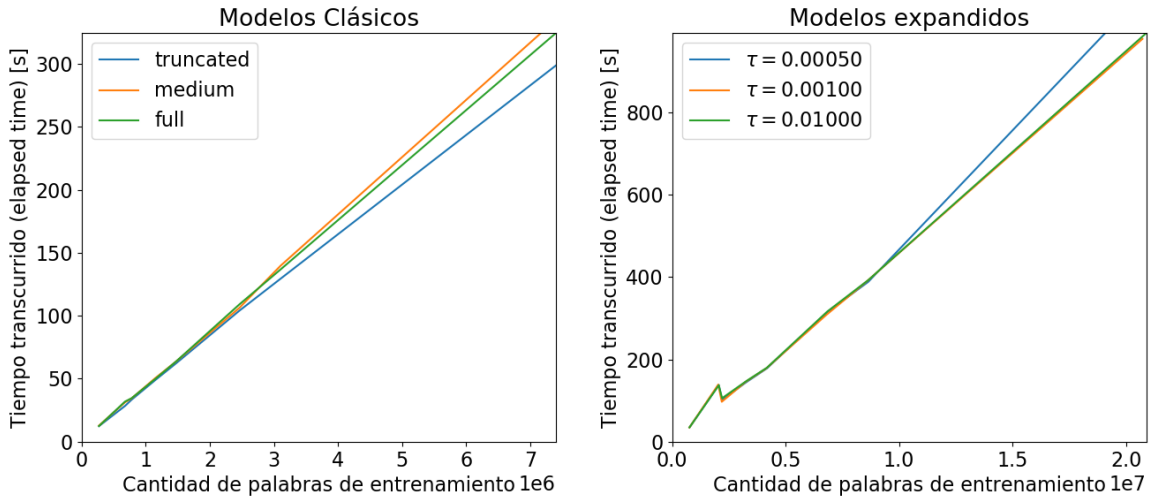


Figura 4.1: Tiempos de entrenamientos para distintos modelos de channeling, en función de la cantidad de palabras o transiciones. A un lado izquierdo se muestra la tendencia del tiempo de entrenamiento para los modelos clásicos (i.e. truncado, medio y máxima resolución); a un lado derecho se detalla la tendencia para los modelos expandidos.

El entrenamiento de un modelo, que posee cerca de 7398317 palabras de entrenamiento, puede alcanzar tiempos cercanos a los 5 minutos para un esquema con *channeling* = 5; un modelo expandido con $\tau = 0{,}00050\,[GHz]$ y un total de 20888617 palabras de entrenamiento tarda cerca de 17 minutos en completar la estimación de parámetros. En otras palabras para un incremento del 282 % en el tamaño del modelo, el tiempo de entrenamiento incrementa en un 339 %. Se aprecia claramente que los modelos con bajo nivel de resolución reducen el

tiempo de entrenamiento de manera considerable, sin embargo, en las siguientes secciones se puede apreciar que estos modelos poseen un bajo poder de predicción sobre datos de observaciones reales.

Otro factor importante a considerar es que el entrenamiento se realiza solo una vez, asumiendo que los parámetros del modelo están bien ajustados. Luego, el resto de la experimentación se realizan sobre los modelos ya entrenados en esta etapa.

4.2. Survey de líneas espectrales

En esta etapa se busca evaluar el poder de clasificación multi-etiqueta. Uno de los problemas clásicos de la espectroscopia es la identificación de líneas espectrales en ventanas espectrales de gran tamaño, donde se observa en un amplio rango de frecuencias para una zona con abundancia de transiciones. El fenómeno de superposición de líneas espectrales, es decir, la identificación de distintas especies moleculares para una misma frecuencia de emisión, es análogo al caso en que una palabra se puede encontrar en más de un tópico. Como L-LDA realiza una estimación sobre distribuciones de probabilidad, el resultado de inferencia sobre nuevos datos entrega el grado de pertenencia sobre todas las etiquetas posibles, así se puede identificar la superposición de transiciones y la presencia de muchas líneas espectrales para una ventana espectral de gran tamaño.

Para evaluar la capacidad del algoritmo L-LDA se reúnen los datos de un *survey* de líneas espectrales realizado por P. Schilke et. al. [28]. La investigación se realiza sobre una región conocida como Orion-KL, una zona con mucha actividad de formación de estrellas. Por esta razón existe una gran abundancia de líneas espectrales, lo que permite evaluar la capacidad del algoritmo propuesto para etiquetar espectros con variedad de transiciones.

El análisis completo cuenta con 1520 transiciones, la tabla 4.2 muestra un resumen de las 5 especies con mayor frecuencia de transiciones, se detalla el total de transiciones y porcentaje de estas transiciones que también están presentes en el modelo. Como se realiza inferencia sobre datos obtenidos de una fuente distinta a los datos con los que se entrenaron los modelos, es necesario aplicar un mapeo de las frecuencias a características del modelo. La

ecuación 3.9 aplica una función de mapeo con un umbral establecido experimentalmente en 0,00020[GHz], así no se realiza mapeo de frecuencias para aquellas transiciones que sobre pasen el umbral. Al realizar el procesamiento con este método se generan espectros con transiciones presentes en modelo y otras que no.

Cuadro 4.2: Cantidad de transiciones de las 5 especies con mayor frecuencia. Datos de testeo recopilados del *survey* de líneas espectrales realizado por P. Schilke et. al. [28]. Se detalla la cantidad de transiciones y el porcentaje de presencia en el modelo.

Especie	Cantidad de transiciones	Presencia
Methanol	457	58,42 %
Dimethyl Ether	420	100 %
Sulfur Dioxide	171	39,8 %
Methyl Formate	145	92,31 %
Ethyl Cyanide	87	73,37 %

Primero se establece una base de comparación con los resultados obtenidos por Mendoza (2018) [22]. Se evalúa el poder de predicción sobre espectros con especies únicas generados con las transiciones de las 5 especies con mayor frecuencia de aparición en el conjunto de datos completo. Para cada una de las especies se obtiene el listado completo de transiciones registradas por P. Schilke, se generan los documentos insertando solo 1 repetición por característica o transición, luego se evalúa el resultado de inferencia realizado por el modelo L-LDA. El procedimiento se repite 5 veces, así se obtiene el valor de Accuracy@1 promediando las repeticiones, es decir, para cada especie se obtiene la proporción de predicciones correctas que se encuentran en el TOP@1. La tabla 4.3 detalla los resultados de la experimentación. Se aprecia que el poder de clasificación es perfecto para las 5 especies más frecuentes del conjunto de datos recopilados por P. Schilke. De esta manera se valida el esquema de experimentación establecido por Mendoza.

En una segunda aproximación, se realizan los mismos experimentos, pero incorporando el operador término-frecuencia tf, que controla la cantidad de repeticiones de una transición en función de la energía E de esta; de esta forma los espectros se encuentran saturados con diversas transiciones de una misma especie, pero cada transición es repetida en proporción

Cuadro 4.3: Accuracy@1 sobre las 5 especies con mayor frecuencia del conjunto de datos recopilado por P. Schilke et. al. [28]. Se establece la inserción de una sola repetición ($tf = 1$) para cada transición.

Especie	Accuracy@1
Methanol	100 %
Dimethyl Ether	100 %
Sulfur Dioxide	100 %
Methyl Formate	100 %
Ethyl Cyanide	100 %

a E. Los resultados se presentan en la tabla 4.4. Se aprecia que los resultados presentan un bajo poder de predicción; para el caso de Dimethyl Ether la clasificación es ideal, pero esto se refuerza debido a la alta presencia palabras en el modelo, con un 100 % de palabras del espectro presentes en el vocabulario. Para el resto de los casos se aprecia un claro problema de sobreajuste producido por la cantidad de repeticiones de cada transición de entrenamiento.

Cuadro 4.4: Accuracy@1 sobre las 5 especies con mayor frecuencia del conjunto de datos recopilado por P. Schilke et. al. [28]. Se establece $tf \propto E$ para cada transición.

Especie	Accuracy@1
Methanol	20 %
Dimethyl Ether	100 %
Sulfur Dioxide	0 %
Methyl Formate	20 %
Ethyl Cyanide	0 %

Como el *survey* de líneas espectrales presenta una gran mezcla de especies moleculares es necesario evaluar el poder de predicción sobre múltiples transiciones de diferentes especies moleculares. Para esto se separó el conjunto de transiciones total en porciones pequeñas para representar distintas ventanas espectrales. Se utilizaron bloques de 30, 100 y 200 transiciones, se define *spw* como la cantidad de transiciones por bloque. La tabla 4.5 muestra un caso de ejemplo para una configuración con $spw = 30$.

Cuadro 4.5: Ventanas espectrales de testeo generadas para ser evaluadas por L-LDA. Se generan 5 ventanas con $spw = 30$, se detalla el rango de cada frecuencia y la cantidad de especies presentes. Datos recopilados por P. Schilke et. al. [28].

Ventana	Rango Frecuencias [GHz]	Cantidad de especies
1	607,17510 – 610,84440	10
2	610,84440 – 612,93160	8
3	612,93190 – 613,85540	3
4	613,85550 – 614,36070	5
5	614,78000 – 623,36350	9

Debido a que Orion-KL es una zona de formación de estrellas y además existe gran abundancia de líneas espectrales, se decide utilizar el modelo *Hot Cores* de resolución máxima (*channeling* = 5); los hiper-parámetros son definidos por el modelo y la cantidad de iteraciones de inferencia se fija en 200. Para cada una de las configuraciones *spw* se evalúa el desempeño del algoritmo en las ventanas espectrales generadas, de esta forma se obtiene la distribución de especies por ventana espectral θ_d. Se utilizan los indicadores P@k y R@k, indicando la proporción de predicciones correctas sobre el TOP k de probabilidades asignadas a cada especie y la proporción de predicciones acertadas sobre el total de especies presentes, respectivamente. En este contexto se identifica una recomendación (predicción) cuando la probabilidad asignada es mayor a la probabilidad más baja. Para cada ventana espectral generada se realizan 10 iteraciones del algoritmo L-LDA para promediar los resultados debido a que el algoritmo presenta variaciones en la estimación de una iteración a otra, esto producto de su naturaleza estocástica. Los resultados de las mediciones P@10 para cada una de las configuraciones *spw* se presentan en la tabla 4.6.

Se aprecia una baja capacidad de predicción para ventanas espectrales con gran cantidad de transiciones. El máximo valor de P@10 se obtiene con la ventana espectral 1 para $spw = 30$ con un puntaje del 40,71 %. Se aprecia que incorporar el operador tf influye considerablemente en la capacidad de predicción del algoritmo L-LDA, esto indica que las repeticiones incorporadas en los entrenamientos de los modelos generan sobreajuste en el modelo. Esto se debe a la presencia de especies moleculares muy dominantes, con muchas repeticiones, que

Cuadro 4.6: Resultados de $P@10$ para los distintos tipos de configuración *spw*. Se detalla el porcentaje de clasificación P@10 para las 5 primeras ventanas generadas con cada configuración *spw*.

	Tamaño ventana		
Ventana	$spw = 30$	$spw = 100$	$spw = 200$
1	40,71 %	36,83 %	0,21 %
2	35,33 %	1,43 %	0,10 %
3	5,33 %	11,94 %	0,09 %
4	16,99 %	0,03 %	0,05 %
5	0,05 %	0,03 %	0,09 %

generan desbalance en la inferencia realizada por el algoritmo. Además se suma el factor de presencia descrito previamente, que indica que las representaciones generadas no se ajustan bien a los datos de entrenamiento.

Los valores de R@10 también son muy bajos por lo que no se presenta el detalle de la experimentación. Se obtiene un máximo valor de 15 % para la ventana espectral 2 con $spw = 30$. Con un mayor nivel de *Recall* estamos asegurando que para un conjunto de transiciones se identifiquen la mayor cantidad de especies moleculares presentes. Los resultados indican que el algoritmo no es capaz de satisfacer la completitud de una búsqueda de este tipo.

De los resultados se identifican algunas situaciones que pueden influir en los resultados:

1. Existen transiciones que no se encuentran registradas en Splatalogue y transiciones cuyas frecuencias difieren en las últimas cifras significativas. Por esta razón el algoritmo no tiene información para inferir respecto a estas transiciones.

2. El algoritmo presenta problemas para clasificar especies moleculares presentes con muchas transiciones. A medida que un documento se comienza a saturar con palabras de un mismo tópico, el algoritmo pierde su poder de clasificación. Esto se produce por el concepto de sobreajuste, que indica que el entrenamiento de modelos presenta un desbalance producto de estas especies con gran cantidad de transiciones.

4.3. Cubos de Datos

En esta sección se presenta una aplicación del algoritmo propuesto para identificar transiciones moleculares individuales en cubos de datos de ALMA. Se desea determinar si el algoritmo L-LDA es capaz de capturar la ocurrencia de transiciones sobre espectros reales utilizando modelos entrenados con datos de Splatalogue. Para la experimentación se utiliza un conjunto de 10 FITS descritos en la tabla 4.7.

Cuadro 4.7: Resumen de los cubos de datos generados por observaciones de ALMA que fueron utilizados para testeo del algoritmo. Se detalla el identificador del proyecto, las dimensiones del cubo, el objeto de interés y la transición presente en el cubo de datos.

Código Proyecto	Dimensiones	Objeto	Transición
2013.1.01268.S	432x432x250	HD163296	Carbon Monoxide (v=0) J=2-1
2013.1.00233.S	512x512x400	IRS43	Formylium (v=0) J=3-2
2012.1.00395.S	450x450x121	Orion IRC2	Hydrogen Isocyanide (v=0) J=3-2
2012.1.00275.S	420x420x330	DM Tau	Carbon Monosulfide (v=0) J=5-4
2011.0.00009.SV	100x100x41	Orion KL	Methanol (v=0) J=8-7
2011.0.00001.SV	100x100x118	TW Hya	Carbon Monoxide (v=0) J=3-2
2011.0.00001.SV	100x100x118	TW Hya	Formylium (v=0) J=5-4
2012.1.00346.S	300x300x200	B355	Carbon Monosulfide (v=0) J=7-6
2012.1.00346.S	300x300x200	B355	Hydrogen Cyanide (v=0) J=4-3
2012.1.00346.S	300x300x200	B355	Formylium (v=0) J=4-3

Para aplicar el algoritmo es necesario obtener la representación descrita en la sección 3.2.3. Se aplica el umbral adaptivo descrito en la ecuación 3.10, estableciendo experimentalmente distintos valores de υ en función de la cantidad de canales del cubo $NAXIS3$. Este valor permite controlar la cantidad de palabras agregadas al documento. Al aplicar este umbral adaptivo se regula que el documento generado no se sature con transiciones que corresponden

a ruido de la observación:

$$\upsilon = \begin{cases} 4{,}0 & \text{si} \quad NAXIS\,3 > 1000 \\ 3{,}0 & \text{si} \quad 200 < NAXIS\,3 < 400 \\ 1{,}5 & \text{etoc} \end{cases} \tag{4.1}$$

Un ejemplo gráfico del umbral descrito se presenta en la figura 4.2 para el espectro procesado desde el cubo de datos que observa DM Tau:

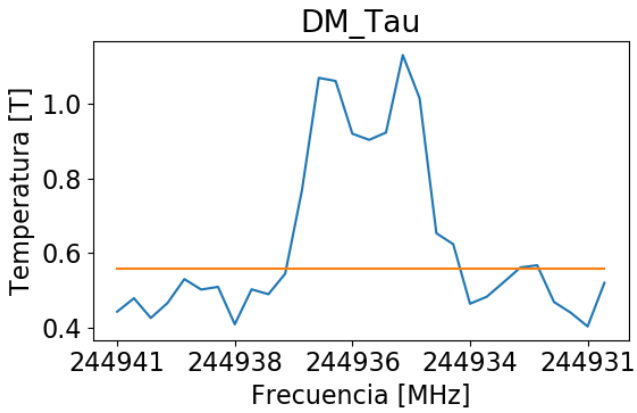


Figura 4.2: Umbral adaptivo establecido como $\upsilon = 2{,}5$ para el objeto de observación DM Tau. Las frecuencias que superan el umbral establecido son consideradas como palabras del documento generado.

Para cada cubo de datos se generan los espectros considerando los distintos niveles de *channeling*, para así evaluar el poder de predicción sobre tres niveles de resolución o *channeling* x distintos. Así se generan 3×10 instancia de testeo para ser evaluadas. Luego de obtener los espectros en formato de documento para cada nivel de resolución y para cada cubo de datos, se evalúa el poder de predicción del algoritmo utilizando L-LDA para estimar las distribuciones de probabilidad de especies por cada instancia. Si la distribución estimada para la especie buscada es la más alta, entonces se establece una clasificación exitosa. Debido a que se presentan distintos modelos L-LDA, basados en el nivel de resolución y en el filtro Splatalogue utilizado, la cantidad de FITS a clasificar difiere en función del modelo utilizado, por esta razón los resultados presentan cierto sesgo hacia esta condición. El Anexo 4.8

detalla la cantidad de FITS de testeo por filtro.

Debido a que L-LDA es un algoritmo de naturaleza estocástica, las predicciones realizadas varían en cada aplicación del algoritmo; por esto las experimentaciones fueron repetidas 10 veces para cada instancia. De esta forma se obtiene el valor promedio de Accuracy@1 correspondiente a cada modelo. Los resultados de la experimentación con los modelos clásicos (*channeling* 0, 2 y 5) se muestran en la figura 4.3. Se aprecia un nulo poder de predicción para los modelos truncados, solo se logra clasificar casos de testeo puntuales. Los modelos de resolución media presentan un mayor poder de predicción, superando incluso a los modelos de máxima resolución para el caso de *Extragalactic*, *Comets* y *Diffuse Clouds*, sin embargo el modelo con mayor Accuracy no supera el 65 %; los modelos con máxima resolución solo fueron mejores para los modelos con mayores cantidades de transiciones y tópicos. En efecto los modelos clásicos no se pudo clasificar Methanol (CH3OH), se aprecia que este componente presentaba gran cantidad de transiciones indicando que el método no presenta dificultad para identificar correctamente transiciones aisladas cuando el tópico es entrenado con gran cantidad de transiciones. Lo anterior indica que se produce sobre ajuste debido a aquellas especies con abundancia de transiciones, como Methanol; esto también se refuerza con el nulo poder predictivo de los modelos con esquema truncado, donde la especies generan muchísimas repeticiones de características aisladas.

Los modelos entrenados con las bandas de ALMA 6 y 7 presentan un muy bajo nivel de generalización en comparación a los modelos entrenados por los filtros de tipo de objeto; por esta razón se muestran los resultados separados. La figura 4.4 presenta el puntaje Accuracy@1 obtenido para la inferencia realizada con estos modelos. Se aprecia que solo los modelos con resolución media logran destacar; para los modelos truncados el poder de predicción es nulo, estos resultados refuerzan la presencia de sobre ajuste descrita anteriormente. Como los modelos truncados consideran solo la parte entera de la frecuencia medida en GHz, al truncar las transiciones de las bandas de alma estamos generando aún más repeticiones de un conjunto acotado de características. El mayor valor de Accuracy@1 fue del 44 % y se obtuvo con la banda 6.

La razón de los bajos niveles de predicción se debe a que el algoritmo está intentando clasificar líneas espectrales de cubos de datos reales con modelos entrenados sobre transiciones

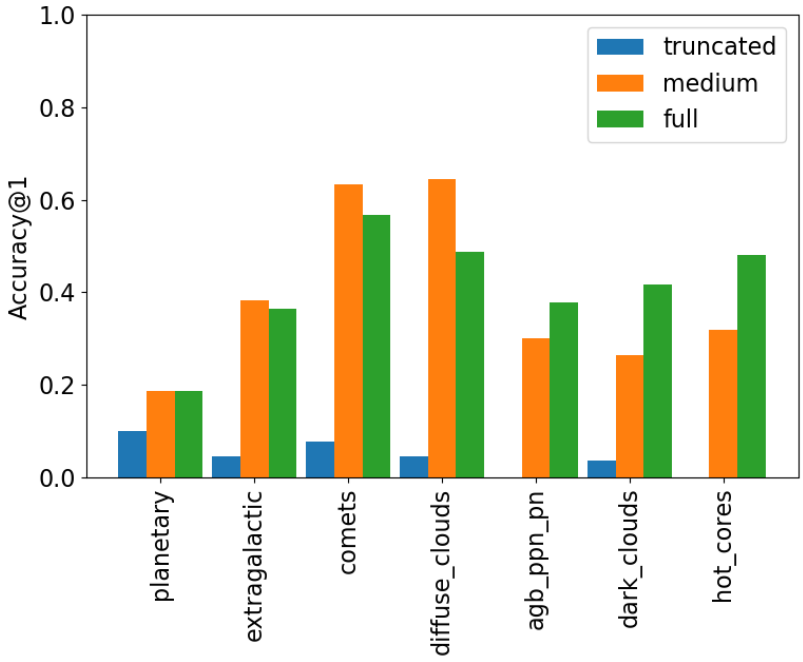


Figura 4.3: Accuracy@1 para modelos clásicos con *channeling* 0, 2 y 5. Los modelos se ordenan en función de la cantidad de palabras de entrenamiento.

puntuales. Con el objetivo de mejorar el poder de predicción se incorporó la expansión de términos descrita en la sección 3.2.2, de esta manera los modelos L-LDA deberían adaptarse mejor para clasificar líneas espectrales aproximadas a campanas de gauss. Al aplicar la expansión de términos al modelo de máxima resolución, el poder de predicción aumenta considerablemente, llegando a obtener un 97,77 % de Accuracy@1 para el modelo*Diffuse Clouds* con parámetros $\tau = 0{,}0100$ y $\chi = 3$. Los resultados de la experimentación con los modelos expandidos se presenta en la figura 4.5. El modelo con soporte $\tau = 0{,}0100$ destaca en todos los tipos de filtros con el mayor poder de predicción, sin embargo los modelos que le preceden en la escala con $\tau = 0{,}0010$ y $\tau = 0{,}0005$ no se alejan mucho del valor máximo correspondiente a cada filtro. El menor valor de Accuracy@1 fue del 56,36 % fue para el caso de *Hot Cores* con $\tau = 0{,}0005$. Se aprecia que L-LDA posee una mayor capacidad de predicción a medida que se considera mayor variedad de transiciones, al utilizar los modelos expandidos con $\tau = 0{,}0100$ se modela una campana de gauss más ancha que permite incorporar una mayor variedad de característica; los modelos con un soporte τ menor presentan menor poder de predicción indicando que la expansión se realiza repitiendo características

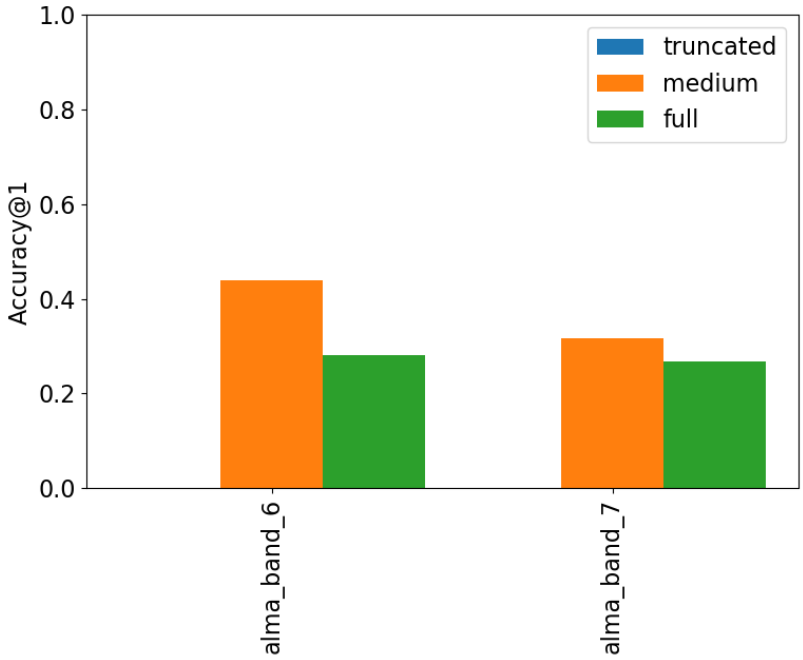


Figura 4.4: Accuracy@1 para modelos Alma Band 6 y 7 con *channeling* 0, 2 y 5. Los modelos se ordenan en función de la cantidad de palabras de entrenamiento.

más que incorporando nuevas.

Para los modelos expandidos de ALMA Banda 6 y 7 los resultados no presentan mejoras en comparación con los modelos no expandidos y por esta razón no se presenta el gráfico respectivo. Para estos modelos el poder de predicción no alcanza el 32 % de Accuracy@1, el esquema de resolución mínima (truncado) logra obtener clasificaciones exitosas, sin embargo el valor de Accuracy@1 no supera el 26 %. Se aprecia que el algoritmo L-LDA presenta problemas para clasificar líneas espectrales utilizando modelos entrenados con muchas especies moleculares.

4.3.1. Tiempos de Inferencia

A pesar de que se obtuvo notables mejorías con los modelos expandidos, la inferencia realizada con estos modelos requiere de mayor tiempo de ejecución. La figura 4.6 presenta el

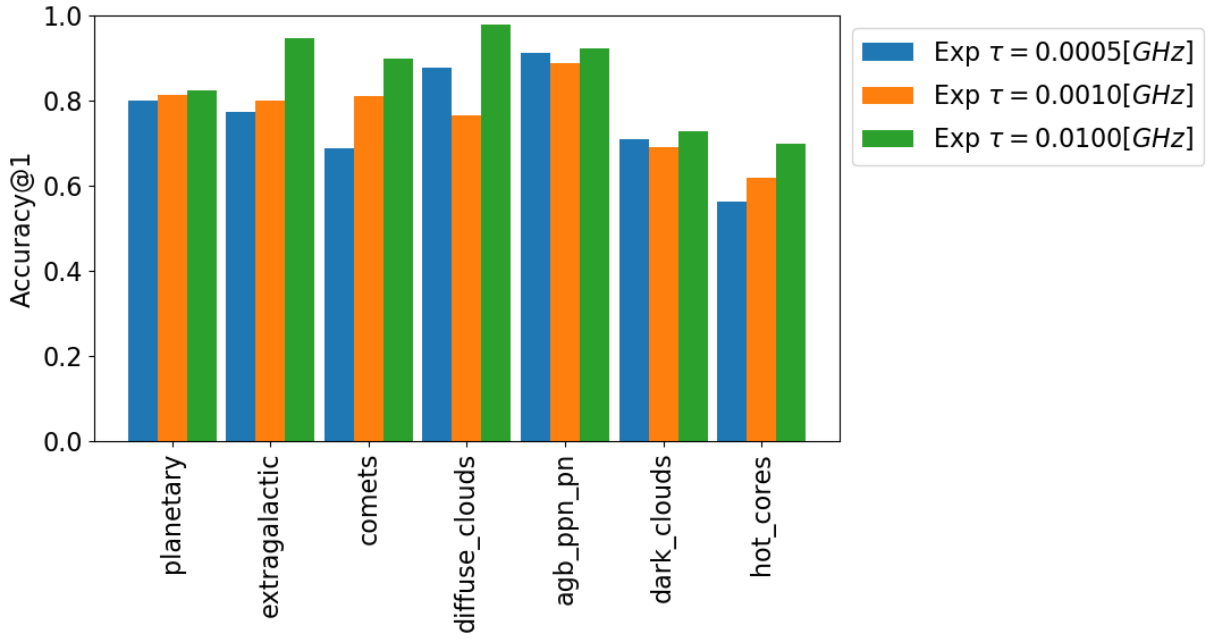


Figura 4.5: Accuracy@1 para modelos expandidos con distintos valores de τ. Se establece la expansión de vecinos como $\chi = 3$. Los modelos se ordenan en función de la cantidad de palabras de entrenamiento.

tiempo transcurrido para realizar inferencia sobre el conjunto total de cubos de datos procesados para cada modelo (ver anexo 4.4).

Se aprecia claramente que los tiempos de inferencia para los modelos expandidos son mucho mayores en comparación a los modelos clásicos, esto se produce debido al gran nivel de resolución que poseen estos modelos. A pesar de que los tiempos incrementan considerablemente, se aprecia un comportamiento lineal de aumento de tiempo en función de la complejidad del modelo.

Utilizar los modelos clásicos asegura que la aplicación del algoritmo de forma rápida, se aprecia que modelo con tamaño cercano a las 70000000 palabras de entrenamiento es capaz de realizar inferencia en aproximadamente 6 segundos para un cubo de datos como IRS43, de tamaño 512x512x400. La inferencia sobre el mismo cubo de datos evaluado con un modelo expandido que posee cerca de 20000000 palabras de entrenamiento puede llegar a demorar cerca de 30 segundos, un valor elevado si se desea implementar el algoritmo para un sistema

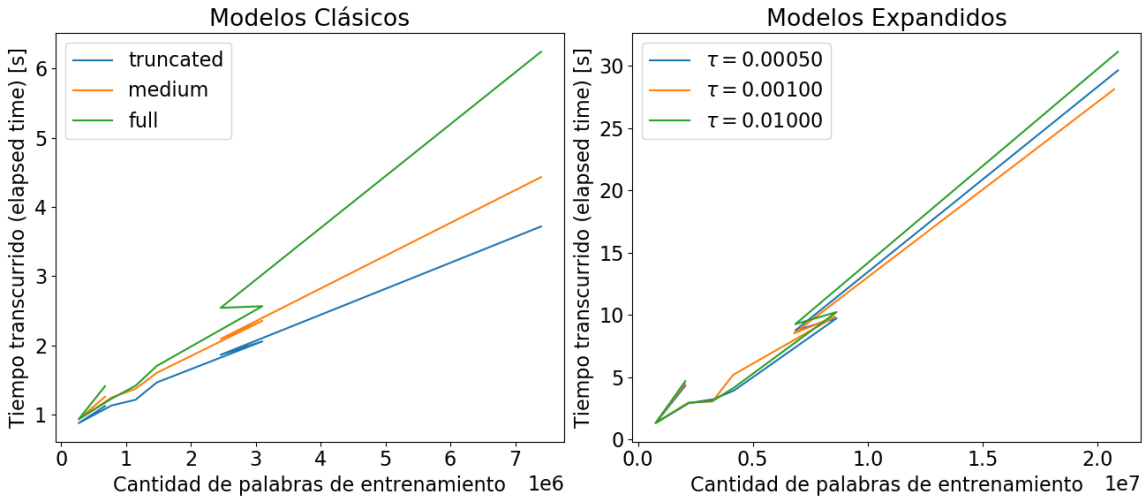


Figura 4.6: Tiempos de inferencia para distintos modelos de channeling, en función de la cantidad de palabras de entrenamiento. A un lado izquierdo se presenta el tiempo de inferencia de los modelos clásicos; a un lado derecho se detalla tiempo para los modelos expandidos. Se detalla el resultado de inferencia para el cubo de datos con mayor volumen, IRS43. Proyecto ALMA 2013.1.00233.S

recomendador en tiempo real. Analizando los casos anteriores, se aprecia que un incremento del 286 % en el tamaño del modelo induce un incremento de cercano al 500 % en el tiempo de inferencia. Al utilizar los modelos expandidos se asegura una predicción mucho más precisa, sin embargo esto induce también un tiempo de ejecución considerablemente mayor.

Conclusiones

El propósito de este trabajo es presentar una aplicación del algoritmo Labeled Latent Dirichlet Allocation orientada a la clasificación automática de líneas espectrales en observaciones astronómicas. Los resultados muestran que la propuesta se puede utilizar para clasificar cubos de datos generados por ALMA, pero la predicción aún no es perfecta. Se requiere mayor análisis en conjunto a expertos para determinar una mejor aplicación del algoritmo en función de los datos con los que se entrena y clasifica.

El poder de predicción multi-etiqueta del algoritmo fue evaluado con datos de un *survey* de líneas espectrales. Los resultados indican que el algoritmo presenta dificultad para capturar la co-ocurrencia de transiciones de distintas especies moleculares. Cuando se construyen los espectros utilizando el operador de término-frecuencia en proporción a la energía de transición se presenta un bajo poder de predicción. Sin embargo, se aprecia que el algoritmo aumenta su capacidad de clasificación cuando se considera solo una repetición por cada transición. Esta indica un claro problema de sobre ajuste en función de los datos de entrenamiento. Se concluye que para generar espectros saturados con especies moleculares a ser clasificadas, se debe considerar sólo el valor de frecuencia de los canales de alta energía en los espectros, sin tomar en cuenta la energía de emisión. Hay que profundizar más con expertos para encontrar una relación más directa entre el operador término-frecuencia y el nivel de energía de las transiciones para así disminuir el sobre ajuste generado.

La experimentación sobre cubos de datos de ALMA indica que a mayor nivel de resolución o *channeling*, mayor es el nivel de accuracy obtenido en la clasificación, alcanzando un 65 % de puntaje con un modelo de resolución máxima y llegando al 98 % de exactitud en la predicción utilizando los modelos expandidos. Al hacer la expansión se aumenta la variedad

de palabras para las que un máximo de intensidad obtiene un acierto, es decir, se presenta mayor diversidad de transiciones relacionadas con una especie, por lo tanto es más fácil para L-LDA generalizar. Se aprecia que existen modelos con mayor poder de predicción que otros, sin embargo todos logran acertar en cubos de datos repetidos; esto indica que existen transiciones que se repiten en los modelos sin importar el tipo de objeto. Así, los modelos son capaces de identificar transiciones en cubos de datos a pesar de que el tipo de objeto no corresponda al modelo con el cual se realiza inferencia.

A pesar de que la propuesta no se puede considerar como una solución definitiva en la clasificación automática de líneas espectrales, si puede presentarse como una herramienta de apoyo para los astrónomos. Herramientas relacionadas con la recuperación de información se pueden implementar utilizando el esquema presentado en este trabajo, de manera de catalogar datos según las especies moleculares con alta probabilidad de presencia. Los resultados respectivos a los tiempos de entrenamiento e inferencia indican que el algoritmo no se puede aplicar para analizar grandes volúmenes de datos en tiempo real debido al considerable tiempo de ejecución del algoritmo. Se requiere un análisis incorporando optimización en el procesamiento de los datos.

Se espera que los instrumentos de observación futuros sean cada vez más complejos y precisos. Por esta razón, a medida que el nivel de resolución de los datos aumente, el nivel de precisión que manejan las observaciones astronómicas será mayor. Así, herramientas de reconocimiento de patrones como la propuesta, podrán identificar líneas espectrales con mayor facilidad debido a que existirá menor diferencia entre los datos de entrenamiento obtenidos de catálogos y los datos de nuevas observación.

Trabajo A Futuro

Evaluar el algoritmo con distintas implementaciones de L-LDA

Para este trabajo se utilizó la implementación JGibbLabeledLDA provista por Myle Ott et. al. La implementación logra cumplir la tarea establecida para clasificar líneas espectrales,

sin embargo, en ciertos casos se obtienen predicciones inesperadas. Por lo anterior, lo más adecuado sería probar el algoritmo con otras implementaciones de L-LDA. la naturaleza del algoritmo de propuesto permite mantener todo el proceso de generación de documentos intacto e incorporar una nueva implementación L-LDA sin problemas.

Realizar optimización de parámetros

En esta experimentación se establecieron todos los parámetros del algoritmo de manera experimental. Los hiper parámetros del modelo y β se calcularon con una fórmula con buenos resultados en la literatura; la cantidad de iteraciones de entrenamiento e inferencia también fueron fijados según estándares; Queda abierta la posibilidad de hacer un estudio de optimización de parámetros para el algoritmo, utilizando técnicas como validación cruzada.

Segmentar instancias de testeo por tipo de objetos

En este trabajo se realizó la experimentación sobre un conjunto de 10 cubos de datos de distintos objetos de interés, provistos por ALMA. Similar al caso de los modelos entrenados con bandas de ALMA, dónde se separó el conjunto de datos de testeo en función de la banda de frecuencias a la que pertenecía cada uno. Por esta razón, incorporar más instancias de testeo y separarlas por tipos de objeto debería aumentar el poder de predicción.

Incorporar conocimiento experto

Es necesario incorporar mayor conocimiento de expertos al modelo para poder clasificar espectros con múltiples líneas espectrales, de manera que las distribuciones estimadas en entrenamiento sean lo más parecidas a las distribuciones presentes en estos conjuntos de datos. Esta es un tarea similar a la expansión de términos realizada en este trabajo, en donde se adapta la distribución de los datos de entrenamiento para representar de mejor manera el tipo de distribución de una línea espectral en astronomía.

Extender la experimentación a otros tipos de datos astronómicos

El algoritmo se puede extender para utilizar distintas fuentes de datos: surveys de líneas espectrales, cubos de datos, espectros en formato FITS e incluso espectros de datos generados sintéticamente. Esto se puede realizar incorporando etapas de procesamiento para cada tipo de datos, con el objetivo de obtener la representación en formato documento estándar que utiliza L-LDA para estimar parámetros.

Clasificar líneas espectrales no identificadas

Otro aspecto que quedó afuera de esta experimentación es la identificación de líneas espectrales desconocidas, es decir, transiciones en dónde aún no se conoce la especie molecular que la origina. LDA es un algoritmo que permite agrupar los datos observados bajo estructuras latentes, sin necesidad de conocer la etiqueta de la estructura previamente; así se puede utilizar este modelo para agrupar aquellas transiciones no conocidas y estudiarlas de manera conjunta.

Optimizar tiempos de ejecución del algoritmo

El proceso del algoritmo que tarda más tiempo corresponde a la estimación de parámetros de cada modelo; sobre todo para los modelos expandidos, que son los que poseen mejor poder de predicción. La etapa de procesamiento del FITS es otra tarea que toma un tiempo significativo para los cubos de datos más grandes por lo que sería interesante incorporar herramientas de procesamiento de FITS eficiente. El tiempo de inferencia de los modelos es marginal en comparación al de entrenamiento, sin embargo, si se desea implementar una herramienta de búsqueda en línea se debería construir una extensión para realizar procesamiento GPU y evaluar los tiempos de ejecución en este ambiente.

Bibliografía

[1] ALMA. http://www.almaobservatory.org/. Accedido: 11 de diciembre de 2019.

[2] ChiVO. https://chivo.cl/. Accedido: 11 de diciembre de 2019.

[3] COSMOS, The SAO Encyclopedia of Astronomy. http://astronomy.swin.edu.au/cosmos/. Accedido: 11 de diciembre de 2019.

[4] LSST. http://lsst.org/. Accedido: 11 de diciembre de 2019.

[5] Splatalogue. http://www.cv.nrao.edu/php/splat/. Accedido: 11 de diciembre de 2019.

[6] M. Araya, G. Candia, R. Gregorio, M. Mendoza, and M. Solar. Indexing data cubes for content-based searches in radio astronomy. *Astronomy and Computing*, 14:23 – 34, 2016.

[7] Mauricio Araya, Mauricio Solar, and Jonathan Antognini. A brief survey on the virtual observatory. *New Astronomy*, 39:46 – 54, 2015.

[8] A. Barrientos and M. Solar. Marchine learning approaches for detection and classification of astrochemical spectral lines. Master's thesis, Universidad Técnica Federico Santa María, Santiago, 2016.

[9] G. Berriman and S. Groom. How will astronomy archives survive the data tsunami. *ACM Queue*, 14:23–34, 2011.

[10] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[11] J. Cernicharo. Laboratory astrophysics and astrochemistry in the herschel/alma era. *European Astronomical Society Publications Series*, 58:251–261, 2012.

[12] C. Comito, P. Schilke, T. G. Phillips, D. C. Lis, F. Motte, and D. Mehringer. A molecular line survey of orion kl in the 350 micron band. *The Astrophysical Journal Supplement Series*, 156(2):127, 2005.

[13] Steven P. Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. *Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond*, chapter 5, pages 129–161. Springer US, Boston, MA, 2012.

[14] Satoshi Eguchi. "superluminal" fits file processing on multiprocessors: Zero time endian conversion technique. *Publications of the Astronomical Society of the Pacific*, 125(927):565, 2013.

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[16] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[17] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[18] Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch. *Handbook of latent semantic analysis*. Psychology Press, 2013.

[19] Frank J. Lovas. Nist recommended rest frequencies for observed interstellar molecular microwave transitions—2002 revision. *Journal of Physical and Chemical Reference Data*, 33(1):177–355, 2004.

[20] S. Maret, P. Hily-Blant, J. Pety, S. Bardeau, and E. Reynier. Weeds: a CLASS extension for the analysis of millimeter and sub-millimeter spectral surveys. , 526:A47, February 2011.

[21] Jon D. Mcauliffe and David M. Blei. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc., 2008.

[22] M. Mendoza, A. Barrientos, M. Solar, and M. Araya. Mixed membership models for source separation of spectral lines. In *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, pages 1–6, July 2017.

[23] Miranda and N. Cabrera. Association rules for spectral lines. Master's thesis, Universidad de Chile, Santiago, 2015.

[24] 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*, 742(1):215 – 227, 2005.

[25] H.M. Pickett, R.L. Poynter, E.A. Cohen, M.L. Delitsky, J.C. Pearson, and H.S.P. Müller. Submilimeter, milimeter, and microwave spectral line catalog. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 60(5):883 – 890, 1998.

[26] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256. Association for Computational Linguistics, 2009.

[27] A. Riveros and K. Pichara. Automatic identification of spectral lines. Master's thesis, Pontifia Universidad Católica de Chile, Santiago, 2016.

[28] P. Schilke, D. J. Benford, T. R. Hunter, D. C. Lis, and T. G. Phillips. A line survey of orion-kl from 607 to 725 ghz. *The Astrophysical Journal Supplement Series*, 132(2):281, 2001.

Anexo

4.4. FITS usados como datos de testing.

Cuadro 4.8: Cantidad de FITS de testeo para cada uno de los filtros de Splatalogue.

Filtro Splatalogue	Cantidad de FITS
Hot Cores	10
AGB/PPN/PN	10
Dark Clouds	10
Extragalactic	10
Comets	9
Diffuse Clouds	9
Planetary	8
Alma Band 6	5
Alma Band 7	6

4.5. Hiper parámetros de los modelos entrenados.

Cuadro 4.9: Hiper-parámetros para los diferentes modelos de datos según el tipo de objeto.

Filtro	x	K	M	α	β
Hot Cores	0	85	9701	0.1	0.0088
	2	85	186902	0.1	0.0005
	5	85	596096	0.1	0,0001
AGB/PPN/PN	0	71	9597	0.7	0,0074
	2	71	84566	0.7	0,0008
	5	71	181268	0.7	0.0004
Dark Clouds	0	58	8830	0.9	0.0066
	2	58	117834	0.9	0.0005
	5	58	227544	0.9	0.0003
Extragalactic	0	36	9032	1.4	0.0040
	2	36	51664	1.4	0.0007
	5	36	65293	1.4	0.0006
Comets	0	30	9349	1.7	0.0032
	2	30	70080	1.7	0.0004
	5	30	98875	1.7	0.0003
Diffuse Clouds	0	24	7684	2.1	0.0031
	2	24	73104	2.1	0.0003
	5	24	120444	2.1	0.0002
Planetary	0	20	6195	2.5	0.0032
	2	20	21973	2.5	0.0009
	5	20	24006	2.5	0.0008

Cuadro 4.10: Hiper-parámetros para los diferentes modelos de datos según la banda de observación 6 y 7 de ALMA.

Filtro	x	K	M	α	β
Alma Band 6	0	133	64	0.4	2.0781
	2	133	6399	0.4	0.0208
	5	133	79617	0.4	0.0017
Alma Band 7	0	132	101	0.4	1.3069
	2	132	9962	0.4	0.0133
	5	132	99040	0.4	0.0013