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

Identificación de Líneas Espectrales Utilizando Descriptores Estadísticos

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

Ciudad, 4 de mayo de 2014

Resumen

La detección y caracterización automática de líneas espectrales es un problema astronómico que aún no ha sido resuelto. El uso de la espectroscopía para describir la composición química de objetos astronómicos a partir de sus líneas de emisión se ve favorecida por nuevas observaciones en regiones de longitudes de onda antes no explorados, que estarán disponibles gracias a proyectos como el Atacama Large Millimeter Array (ALMA). Con el uso de datos simulados basados en las observaciones que se tendrán a partir de ALMA, se propone la extracción de descriptores estadísticos de líneas emisión pertenecientes a estructuras astronómicas para su uso en un modelo de clasificación que lidie con la incerteza de las mediciones. Esto significa que dicho modelo debe ser capaz tanto de identificar a que moléculas probablemente corresponden las líneas de emisión, así como entregar una probabilidad de pertenencia. A partir de este algoritmo, los astrónomos podrán determinar la composición molecular probable de objetos astronómicos.

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

1. Resumen Ejecutivo

El contexto de este informe es el de dar a conocer en qué estado se encuentra el área de investigación de la identificación de líneas espectroscópicas desde la perspectiva del aprendizaje de máquina y la minería de datos, en el marco del desarrollo de un proyecto colaborativo para la iniciativa del Observatorio Virtual Chileno (ChiVO).

Para el desarrollo del algoritmo se utilizarán datos simulados que emulan las características de las futuras mediciones del proyecto Atacama Large Millimeter Array (ALMA), con los cuales se aplicarán técnicas de análisis de datos para la detección, caracterización y clasificación de líneas espectrales.

En este documento se detalla la proposición de un algoritmo de detección de líneas espectroscópicas utilizando descriptores estadístcos. Esto se hará a partir de la búsqueda de patrones entre líneas de emisión provenientes de estos cubos simulados, los cuales abarcan el rango de frecuencias del proyecto ALMA.

Con un conjunto simulado de espectros se pretende identificar líneas a partir de los parámetros con los que fueron simuladas dichas líneas espectrales, con lo que se entregaría la distribución de pertenencia de estas líneas a algun tipo de molécula en particular.

La particularidad del algoritmo de clasificación propuesto radica en la incorporación del factor incerteza de las mediciones, donde tanto los datos de input como el resultado del algoritmo consistirán en distribuciones de probabilidad en vez de datos específicos, con lo cual se capturará toda la riqueza disponible de los espectros a la hora de predecir líneas de emisión.

A lo largo del documento se detalla el estado del arte actual de la clasificación, de las distintas aproximaciones a los algoritmos de clasificación que incorpotan incerteza y finalmente se detallarán las especificaciones de requerimientos del algorito a desarrollar.

Índice

2. Metodología de Trabajo

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

El desarrollo de este proyecto de investigación surge de la potencial capacidad del aprendizaje de máquina y de la minería de datos de ser utilizados como herramienta para resolver complejos problemas astronómicos que involucren una gran cantidad de datos a ser procesados.

Con proyectos como el Atacama Large Millimeter Array (ALMA), la cantidad de datos de mediciones astronómicas que se tendrán a disposición crecerá exponencialmente, por lo que una herramienta que automatice el procesamiento de dichos datos y entregue información útil para los astrónomos es de gran utilidad.

2.1. Identificación de problemas a resolver

En una etapa inicial, se efectuaron una serie de reuniones administrativas con el conjunto de universidades que colaboran en el proyecto ChiVO. En estas reuniones se buscaba establecer como abordar y en qué se centraría cada universidad a la hora de aplicar técnicas de minería de datos a la información que ALMA proporcionará.

Con la ayuda de astrónomos fue posible realizar una propuesta que satisfaciese los requerimientos de un problema astronómico, donde cada proyecto de las univesidades asociadas al área de minería de datos resolvería dicho problema con una aproximación diferente.

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

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

Con esta información se contaría con una gran cantidad de líneas espectroscópicas para cada punto espacial, por lo que surge la detección de líneas espectrales como un problema de interés común para aplicar técnicas de aprendizaje de máquina.

Por el lado del aprendizaje de máquina, se contará con suficientes datos para entrenar un algoritmo

de clasificación automática, y por parte de los astrónomos, se podrá contar con una herramienta que automatice la tarea de identificar líneas de emisión en un conjunto de espectros.

Luego de definir el problema, era necesario determinar un set de datos con el cual comenzar a desarrollar el algoritmo. Dada la cantidad necesaria de espectros para entrenar un modelo representativo y que generalice adecuadamente, y como actualmente no se cuenta con mediciones adecuadas del radiotelescopio ALMA, se optó por utilizar una simulación de espectros.

2.2. Simulación de datos de entrenamiento

Para el desarrollo del algoritmo de identificación de líneas espectrales se utilizará el servicio web de datos simulados que proporcionará ChiVo, el cual que se desarrolla en paralelo a este proyecto como parte de las herramientas que podrán utilizar los astrónomos como servicio web.

La simulación permitirá generar un set de entrenamiento para el uso del clasificador supervisado que se propone en este proyecto. Al ser este servicio de simulación un input importante para el desarrollo del algoritmo, se han realizado una serie de reuniones para trabajar conjuntamente y obtener un set de datos simulados.

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

Para que el algoritmo sea posible de entrenar con los datos simulados, se deberá incluir en la metadata de los cubos las líneas de absorción presentes en los espectros, así como su distribución de estar realmente donde la simulación ha instanciado dicha línea.

Así, el algoritmo será capaz de mapear cada distribución con la línea espectroscópica respectiva y con la adición de características tanto de la frecuencia donde probable se encuentra, así como descriptores de la forma de la línea espectroscópica y correlación de las mediciones de una línea en particular, encontrar patrones para realizar el proceso de clasificación.

2.3. Especificación de Requerimientos

El método de trabajo consiste en realizar reuniones mensuales con todas las universidades del proyecto para analizar avances y la entrega de hitos con el cumplimiento secuencial de los diferentes requerimientos del proyecto del observatorio virtual.

Se definieron estándares para todos los proyectos involucrados con el fin de entregar un producto final congruente con la idea de generar un paquete de herramientas astronómicas de ChIVO disponible para la comunidad astronómica. Por lo mismo, se definió el uso del lenguaje de programación Python por su fácil integración con el software CASA como lenguaje común entre todos los servicios.

El algoritmo de detección de líneas espectroscópicas será un servicio web. Como imput se recibirá una imágen FITS de una estructura de un objeto astronómico simulado con el servicio web disponible en el proyecto ChiVO. Como output se entregará un archivo con las líneas de emisión detectadas y la distribución de posibles moléculas a las que podrían pertenecer dichas líneas, con la probabilidad asociada a cada predicción realizada.

3. Estado del Arte

3.1. Descripción General

El área de machine learning y minería de datos tiene mucho potencial para ser aplicado en problemas que involucran gran cantidad de datos, como lo es el área de la radio-astronomía y la identificación de líneas de emisión.

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

La falta de escalabilidad de técnicas que no sean automatizadas, y lo poco práctico que resultan dichos método para grandes cantidades de datos [13], añadido a la dificultad a la hora de predecir nuevas coincidencias entre frecuencias y moléculas dada por las superposiciones de líneas, ha impulsado a los astrónomos a buscar la automatización de esta tarea.

El problema de mezclas de líneas (blending) y superposiciones (beams) son producto de tanto ruido como la falta de sensibilidad para distinguir entre dos líneas en frecuencias cercanas. Lo anterior también puede producir peaks dobles en ciertas líneas [3].

Un problema importante a la hora de identificar frecuencias subyace en líneas ópticamente delgadas, que tienden a dar resultados incorrectos. Usualmente, el uso de líneas de isótropos para su corrección resulta en un proceso costoso en tiempo y por lo mismo no es apto para datos masivos [13].

Nummelin et al. [11] propone el uso de un ajuste manual de las líneas a una forma arbitraria dada por una gaussiana, obteniendo por cada línea su frecuencia observada, el peak en el brillo de tempratura y el ancho de la velocidad (ancho total a media altura), para así proceder con la identificación de la línea al asociarla con una molécula en cierto estado de energía.

Para la identificación de líneas considerando las relaciones entre brillo de temperatura en un mismo espectro, es necesario asumir temperatura y origen homogéneo, dado que la diferencia de temperatura cambia la relacion en serie de intensidades de líneas hiperfinas [12].

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

Es posible detectar patrones en las líneas que corresponden a la misma molécula a partir de intensidad relativa considerando que existe una razón entre diferencias de velocidad que es constante para un conjun-

to de líneas de emisión. Esto permite buscar patrones no tan solo de manera individual, sino que a través del análisis manual de series de líneas que se asocian a una misma molécula o átomo en sus diferentes estados energéticos.

Las técnicas anteriormente descritas no son escalables al no ser procesos automatizados y depender de análisis o ajustes manuales que con la inminente llegada de enormes cantidades de datos provenientes de instrumentos como ALMA, dejan de ser aplicables. Por esto es necesario buscar algoritmos de clafication que deleguen la tarea de identificar y clasificar líneas espectrales.

3.2. Soluciones Actuales

Los esfuerzos para desarrollar una herramienta automática de detección de líneas actualmente se limitan a herramientas semi-automáticas que utilizan como base complejos modelos físicos y químicos para la clasificación de líneas. Dichos modelos son aplicados en solo una medición en un espectro determinado, por lo que no se consideran las corelaciones existentes entre distintas mediciones de espectros para un mismo objeto. [19].

Estas herramientas hacen uso de la colección de bases de datos astronómicas que contienen información sobre líneas espectrales de moléculas y sus frecuencias teóricas de laboratorio, las que están disponibles públicamente en catálogos como Splatalogue [16, 15].

3.2.1. Clasificación automática determinista

Precedentes del uso de aprendizaje de máquina para clasificación supervisada de objetos astronómicos son el uso de ajustes de modelos de regresión y de indicadores de autocorrelación para predecir el tipo de objeto o realizar regresiones para alguna variable en particular, lo que ha dado buenos resultados en diferentes áreas de la astronomía.

La extracción de descriptores estadísticos de las series de tiempo ha permitido la clasificación de objetos estelares como estrellas en diferentes subclases según la variación de su brillo, mediciones que quedan plasmadas en las llamadas curvas de luz [5, 17, 18, 9].

La clasificación supervisada corresponde a una técnica donde el modelo predictor toma como input un set de entrenamiento compuesto por variables, que consisten en descriptores de los objetos a clasificar, y obtiene una etiqueta del objeto según sus características, lo que se llama la clase del objeto. En el caso de la detección de líneas las variables serían descriptores estadísticos y la clase o etiqueta a predecir sería la molécula a la que pertenece dicha línea.

Técnicas comunes para la clasificación supervisada en Machine Learning son la mezcla de Gaussianas [1], los árboles de decisión [14], Naive Bayes [6], Redes Neuronales, Support Vector Machines [4], y Random Forest [2]. Estos métodos son modelos de aprendizaje de máquina que aprenden a predecir variables categóricas a partir de un set de otras variables de cualquier tipo.

3.2.2. Clasificación automática probabilística

Los modelos de clasificación mencionados anteriormente pueden predecir la molécula a la cual pertenece una línea, o un conjunto de estas, dadas las características de descritas según la representación que sus descriptores estadísticos entreguen.

Sin embargo, existen limitaciones e incertezas en las mediciones dadas por fenómenos como el blending que dificultarían la clasificación al tener que utilizar solo una etiqueta en particular, lo que se conoce como clasificación determinista.

Existen diversos modelos de clasificación que consideran la incerteza de los descriptores estadísticos, entregando una distribución de probabilidad de las posibles clases a las que el objeto puede pertenecer. Los principales son los árboles de decisión y los los Fuzzy Trees [20, 8].

Las Belief Trees son un tipo especial de fuzzy trees que integra la teoría de las belief functions, donde se agrega una capa de abstracción a la teoría probabilística y se trabaja con la teoría de conjuntos de set de variables probabilísticas [7].

Sin embargo, estos modelos no entregan etiquetas o clases a predecir que reflejen la incerteza de las mediciones a partir de un input con incerteza, por lo que actualmente no existe un clasificador que sea capaz de tomar las mediciones con su incerteza asociada, y entregar como resultado una distribución probabilística de posibles etiquetas a las que pertenece el objeto a clasificar, lo que se llama clasificación probabilística.

Referencias

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

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

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

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

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

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

[7] Z Elouedi, K Mellouli, and P Smets. Belief Desicion Trees: theoretical foundations. *Machine learning*, pages 91–124, 2000.

[8] I Jenhani, N Ben Amor, and Z Elouedi. Decision trees as possibilistic classifiers. *Machine learning*, 2008.

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

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

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

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

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

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

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

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

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

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

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

[20] P Vannoorenberghe and T Denoeux. Handling uncertain labels in multiclass problems using belief decision trees. *Machine learning*, 2002.

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

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