Análisis de Datos Astronómicos

Karim Pichara
Profesor Asistente, DCC Pontificia Universidad Católica de Chile. Investigador del Centro de Astro-Ingeniería y del grupo de Biomedicina de la Pontificia Universidad Católica de Chile. Doctor en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile (2010). Posdoctorante en el laboratorio "Time Series Center" del Centro de Astrofísica de la Universidad de Harvard (2011-2013).
kpb@ing.puc.cl

Rodolfo Angeloni
Investigador posdoctoral, DAA Pontificia Universidad Católica de Chile. Doctor en Astronomía (2009), Universitá di Padova, Italia . Investigador Responsable del Proyecto FONDECYT N. 3100029 "Topics in Stellar Variability: from VISTA to ALMA".
rangelon@astro.puc.cl

Susana Eyheramendy
Profesora asistente, Depto. de Estadística Facultad de Matemáticas, Pontificia Universidad Católica de Chile. PhD Depto. de Estadística Universidad de Rutgers, EE.UU; posdoc Universidad de Oxford y Ludwig-Maximilian Universitat/Instituto de Epidemiología del Helmholtz Zentrum Munich, Alemania. Su investigación se basa en el análisis y desarrollo de métodos en estudios genéticos de asociación y en aplicaciones de métodos de minería de datos a problemas astronómicos.
susana@mat.puc.cl

En el Centro de Astro-Ingeniería de la Pontificia Universidad Católica de Chile, un grupo de científicos conformado por los profesores Márcio Catelan, Andrés Jordán y Rodolfo Angeloni de Astronomía; Susana Eyheramendy de Estadística; Karim Pichara de Ingeniería en Ciencia de la Computación, y el alumno de Ingeniería Cristóbal Berger, se dedican al desarrollo de herramientas inteligentes para el análisis de Datos Astronómicos. Durante los últimos años, ha existido un creciente interés en aplicaciones de inteligencia artificial (Russel and Norvig (2010)) y aprendizaje de máquina (Mitchel (1997)) para la investigación astronómica debido al gran desarrollo tecnológico de los telescopios, cada vez capaces de generar una mayor cantidad de información imposible de ser analizada en su totalidad por humanos. Por ejemplo, el próximo telescopio LSST ("Large Synoptic Survey Telescope"[1]) tendrá la labor de producir durante diez años alrededor de 30 Terabytes diarios de información proveniente del Universo, esto corresponde a varios billones de objetos, cada uno observado en alrededor de 1.000 instantes distintos de tiempo. El proyecto "Vista Variables in the Via Lactea (VVV) ESO Public Survey"[2], escaneará la Vía Láctea arrojando mediciones en la banda infrarroja de más de diez mil millones de objetos en el espacio. Estos desarrollos impulsan nuevas necesidades científicas: a mayor cantidad de información disponible, mayor es la necesidad de nuevas tecnologías para el análisis de estos datos.

Una de las tareas más importantes en el análisis de datos del espacio es la clasificación automática de objetos estelares. Existe hoy gran interés en desarrollar modelos de

1 http://www.lsst.org/lsst/
2 http://mwm.astro.puc.cl/mw/index.php/Main_Page

Figura 1

Segundo lanzamiento de la imagen del VST, probablemente es el mejor retrato del cúmulo globular Omega Centauri que alguna vez se haya obtenido. Omega Centauri, en la constelación de Centaurus es el cúmulo globular más grande del cielo.

Cuando sea observado con el telescopio infrarrojo VISTA, nuestro grupo logrará obtener curvas de luz para un número importante de diferentes tipos de estrellas variables. Estas observaciones constituirán una fracción importante del conjunto de entrenamiento que estamos construyendo en el proyecto "VVV Templates".

aprendizaje de máquina capaces de aprender a clasificar automáticamente estos objetos a partir de bases de datos previamente rotuladas por astrónomos (Debosscher et al. (2007), Dubath et al. (2011), Richards et al. (2011), Kim et al. (2011)). Estos sistemas de clasificación deben considerar desde el preprocesamiento de los datos hasta la generación del modelo capaz de clasificar automáticamente los objetos.

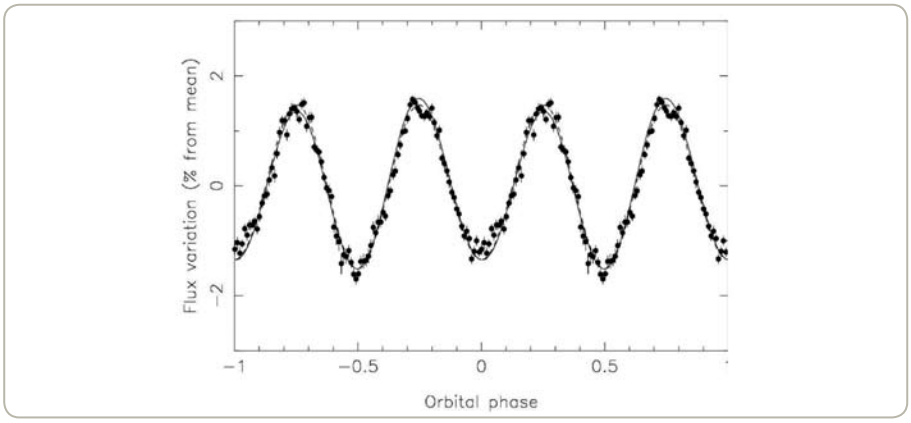
Dado que la mayoría de los proyectos observacionales como LSST y VVV incluyen observar estrellas variables (estrellas que muestran una variación en su brillo en función del tiempo. Ver catálogo de las distintas clases en http://www.sai.msu.su/gcvs/gcvs/iii/vartype.txt), es natural enfocar los esfuerzos en el análisis de series de tiempo o curvas de luz (gráfico que se obtiene de la variación del brillo en función del tiempo, Figuras 2, 3 y 5), este análisis busca representar en forma compacta una curva de luz de tal manera de simplificar la información que recibe un algoritmo de clasificación.

En la literatura existe una amplia gama de modelos de análisis de series de tiempo (Percival et al. (2003), Mills et al. (1990), Bloomfield (1976), Hamilton (1994)). La principal ventaja de usar modelos para analizar las series de tiempo es poder extraer características propias de la forma de cada curva de luz, de tal modo de obtener información útil para que los algoritmos de clasificación automática puedan desempeñar su labor usando como principal información estas características obtenidas del análisis de cada serie de tiempo. Existen numerosas técnicas para modelar curvas de luz (Lomb (1976), Scargle (1982), Ponman (1981), Kurtz (1985)). Estos modelos estiman los parámetros y frecuencias de un modelo armónico sobre la forma de la curva de luz, de tal modo de usar los parámetros encontrados como descriptores de cada curva.

Consideremos como $y(t)$ la intensidad de luz observada en un instante t, sea $\hat{y}(t)=a+bt$ una estimación lineal de $y(t)$ y sea $r(t)=y(t)-\hat{y}(t)$. Iteramos entre los siguientes pasos:

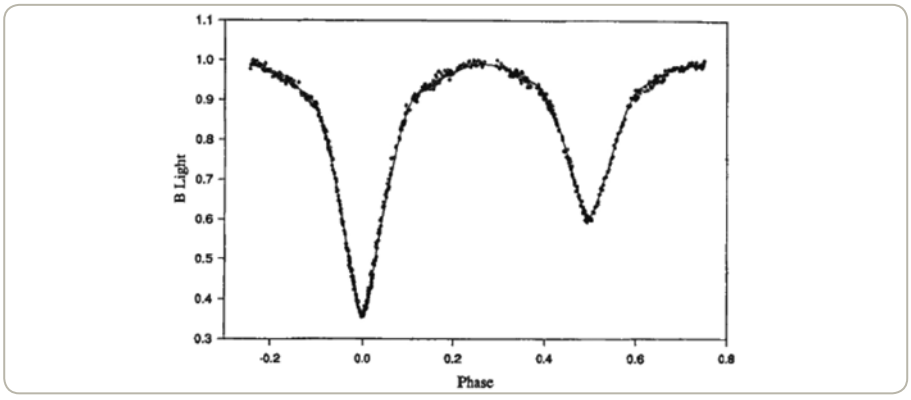
1. Realizar un análisis de Fourier para $r(t)$ con el objetivo de determinar cualquier periodicidad que podría existir usando el método Lomb-Scargle(Lomb (1976), Scargle (1982)). Una vez calculado el periodograma de Lomb-Scargle se selecciona el valor máximo. La frecuencia correspondiente

Figura 2



Curva de luz de KPD1930+2752 después de remover la señal debido a las pulsaciones de Billères et al. (2000) con un modelo de curva de luz (línea sólida) para la variabilidad elipsoidal asumiendo una inclinación de 90°.

Figura 3



Curva de luz de TT Aurigae observada por Wachmann, Popper, y Clausen (1986) y modelada por Terrell (1991).

Una de las tareas más importantes en el análisis de datos del espacio es la clasificación automática de objetos estelares. Existe hoy gran interés en desarrollar modelos de aprendizaje de máquina capaces de aprender a clasificar automáticamente estos objetos a partir de bases de datos previamente rotuladas por astrónomos.

f se usa para encontrar los parámetros de la siguiente función armónica, usando un método de mínimos cuadrados:

$$\hat{z}(t) = \sum_{j=1}^{m} (a_j \sin 2\pi f jt + b_j \cos 2\pi f jt) + b_0$$

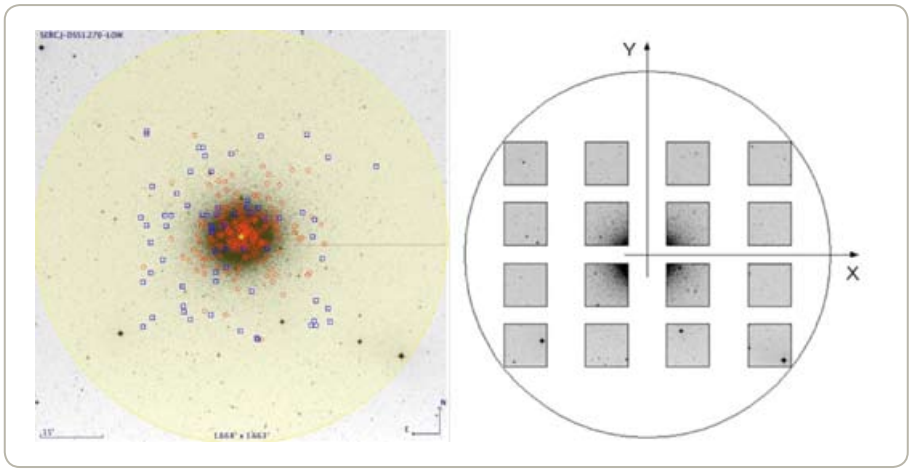
2. Actualizar $r(t)=r(t)-\hat{z}(t)$

En palabras, primero se resta la tendencia lineal de la serie de tiempo fotométrica, luego usando el periodograma de Lomb-Scargle identificamos el peak más alto y usamos la frecuencia correspondiente para ajustar un modelo armónico con m componentes. Esta nueva curva, junto con la estimación lineal son restadas desde la serie de tiempo y se busca una nueva frecuencia en el residuo usando el periodograma de Lomb-Scargle, la nueva frecuencia se usa para ajustar nuevamente el modelo armónico. Este proceso continúa hasta que se encuentran n frecuencias y se estiman n modelos armónicos con m componentes. Finalmente, las n frecuencias se usan para realizar el mejor ajuste a la curva de luz original:

$$\hat{y}(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij} \sin 2\pi f_i jt + b_{ij} \cos 2\pi f_i jt) + a + bt$$

Las frecuencias f_i junto con los parámetros de Fourier a_{ij} y b_{ij}, constituyen el conjunto de parámetros con los cuales podemos representar las curvas de luz.

Figura 4



Próximas observaciones VISTA del cúmulo globular omega Cen. Panel izquierdo: un ejemplo de distribución a lo largo del cluster: los círculos rojos marcan la posición de estrellas RR Lyrae conocidas, los cuadrados azules marcan las posiciones de estrellas binarias eclipsantes, conocidas. Panel derecho: 16 detectores de VIRCAM@VISTA, con omega Cen en el centro del plano focal.

Uno de los principales problemas de esta representación es que los parámetros no son invariantes a traslaciones en el tiempo. En otras palabras, si de la misma estrella tenemos dos curvas de luz observadas para las cuales no coincide el instante de tiempo inicial, estas dos curvas de luz tendrán un conjunto distinto de parámetros representando la misma estrella. Para lidiar con este problema transformamos los coeficientes de Fourier en un conjunto de amplitudes A_{ij} y fases PH_{ij} como sigue:

$$A_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2},$$

$$PH'_{ij} = \arctan(\sin(PH_{ij}), \cos(PH_{ij}))$$

donde:

$$PH_{ij} = \arctan(b_{ij}, a_{ij}) - \frac{jf_i}{f_1} \arctan(b_{11}, a_{11})$$

Notar que PH_{11} es elegido como la referencia y es igual a cero, además $PH`_{ij}$ toma valores en el intervalo $]-\pi,\pi]$

Una vez que tenemos una representación paramétrica para las curvas de luz, podemos utilizar modelos de clasificación para aprender a identificar cada tipo de estrella.

Existen numerosos algoritmos de clasificación automática en la literatura de Aprendizaje de Máquina (Mitchel (1997), Bishop (2006)). Los algoritmos de clasificación están basados principalmente en modelos matemáticos para encontrar espacios de separación entre las distintas clases de objetos, entre los algoritmos más nombrados están los "árboles de decisión" (Quinlan (1986)), "Support Vector Machines" (Cortes & Vapnik 1995), el clasificador "Naive Bayes" (Mitchel (1997)), el "clasificador de vecinos cercanos" (Mitchel (1997)), etc. El proceso de aprendizaje de estos algoritmos consta en utilizar un conjunto de instancias para entrenar (conjunto de entrenamiento) donde el algoritmo busca separar entre los elementos de distintas clases para luego probar el rendimiento del modelo de clasificación en un conjunto

de instancias que no fueron usadas en el proceso de entrenamiento (conjunto de evaluación). Es importante a la hora de iniciar el entrenamiento de un modelo de clasificación considerar que el objetivo final es que el modelo clasifique con un alto rendimiento las instancias del conjunto de evaluación, de tal modo de probar que el modelo es capaz de clasificar correctamente instancias nuevas, no procesadas durante el entrenamiento, eso asegura una buena capacidad de generalización del modelo de aprendizaje.

Al momento de lidiar con Datos Astronómicos aparecen muchas limitaciones que hacen más difícil el proceso de aprendizaje de clasificadores. Una de estas limitaciones corresponde al gran costo de obtener datos etiquetados para formar el conjunto de entrenamiento. Dado que los telescopios no arrojan información sobre el tipo de objeto que inspeccionan, sino que sólo información sobre algunas de sus características, es necesario que los astrónomos manualmente se dediquen a etiquetar estos datos de tal manera que un computador pueda iniciar el proceso de aprendizaje.

Dado también que algunos tipos de Datos Astronómicos existen hace muchos años, hoy gran parte de ellos se encuentran etiquetados y disponibles para la comunidad científica, éste es el caso de los datos ópticos. Lamentablemente existen otros tipos de datos que no están etiquetados por la comunidad astronómica, por ejemplo el VVV es la primera inspección de variabilidad estelar en la banda infrarroja, en este caso nuestro grupo debe lidiar con la construcción de una base de datos etiquetada de variabilidad estelar en la banda infrarroja. Así el principal propósito del proyecto "VVV Templates"[3] se traduce en construir un conjunto de entrenamiento en la banda infrarroja para los clasificadores automáticos, que hasta ahora sólo han sido utilizados en datos ópticos.

Parte del proyecto "VVV Templates" comprende observar el cúmulo globular Omega Cen (Figuras 1 y 4). Este cúmulo contiene varios millones de estrellas, entre ellas varios cientos de estrellas variables de diferentes tipos y nos permitirá obtener con una serie de observaciones una fracción importante de los datos que se necesitan para construir los "templates" de curvas de luz que se esperan obtener de este proyecto.

Con todos los recursos que se necesitan para obtener estos "templates" nace la necesidad de implementar modelos de clasificación con un nivel mayor de inteligencia, capaces de seleccionar eficientemente sólo los objetos más informativos de tal modo de aprender a clasificar con la menor cantidad de información posible, de tal modo de ahorrar recursos valiosos como el tiempo dedicado a las observaciones al telescopio. Este proceso de seleccionar instancias específicas es conocido en la literatura del aprendizaje de máquina como "active learning" o aprendizaje activo (Roy et al. (2001), Cebron et al. (2008), Baram et al. (2004)). Los modelos de aprendizaje activo van seleccionando en cada iteración la instancia más apropiada para el aprendizaje a partir de un conjunto de instancias candidatas, luego solicita a algún ente experto (en este caso el astrónomo) que clasifique la instancia previamente seleccionada, para luego incluir esta información en el conjunto de entrenamiento y refinar el clasificador y el modelo selector de instancias.

Más específicamente, considere un conjunto de instancias (curvas de luz) descritas por d parámetros $X=\{x_1,...,x_n\}$, donde $x_i \in R^d\, i \in [1,..,n]$ y un conjunto de C posibles clases de estrellas $Y=\{y_1,...,y_c\}$, donde cada x_i pertenece a una clase y_j, $j \in [1,..,C]$. Considere el conjunto $U \in X$ de curvas de

Figura 5

La "Nebulosa del Cangrejo del Sur", uno de los mejores ejemplos de estrella variable de tipo simbiótico en sus últimas fases de evolución. Esta imagen fue obtenida utilizando el telescopio espacial Hubble. Autores: Romano Corradi, Mario Livio, Ulisse Munari, Hugo Schwarz y NASA.

3 http://www.vvvtemplates.org/

luz no clasificadas y el conjunto $L \in XxY$ de curvas previamente etiquetadas (la clase de cada elemento en *L* es conocida).

El proceso de aprendizaje activo consiste en estratégicamente seleccionar curvas de luz desde el conjunto *U* para que sean etiquetadas por un astrónomo y luego agregadas al conjunto *L*. Cada vez que *L* cambia se actualiza un clasificador cuyo conjunto de entrenamiento es *L*. Principalmente la idea es seleccionar las curvas de luz que más aportan en el entrenamiento del clasificador. Sea *P(y|x)* la distribución (desconocida) que corresponde a la probabilidad de que un dato *x* pertenezca a la clase *y*, sea *P(x)* la distribución de probabilidades marginal sobre las instancias. Sea $\hat{P}_D(y|x)$ la distribución de probabilidades que el modelo clasificador debe aprender del conjunto de entrenamiento *D*. El error esperado de la clasificación es:

$$E_{\hat{P}_D} = \int_x Ls(P(x|y), \hat{P}_D(y|x))P(x)$$

Donde *Ls* es una función que mide el grado de pérdida o diferencia entre la distribución estimada y la distribución real:

$$Ls = \sum_y P(y|x)log(\hat{P}_D(y|x)))$$

El algoritmo de aprendizaje activo seleccionará la instancia x_i tal que al añadirla al conjunto *L* el clasificador entrenado con el conjunto resultante $L^* = L + (x_i, y_i)$ obtiene el menor error comparado con todas las otras instancias candidatas, es decir:

$$\forall(x,y)E_{\hat{P}_{L^*}} < E_{\hat{P}_L}$$

Lamentablemente la distribución real *P(y|x)* es desconocida, por lo tanto realizamos una estimación del error en base al valor esperado de la clasificación de cada instancia, usando el clasificador que se tiene hasta el momento evaluado sobre el conjunto de testeo, así el error estimado se calcula como:

$$\tilde{E}_{\hat{P}_{L^*}} = \frac{1}{|T|}\sum_x \sum_y \hat{P}_{L^*}(y|x)log(\hat{P}_{L^*}(y|x))$$

Donde $|T|$ es el número de elementos en el conjunto de testeo.

En palabras simples, el modelo va a elegir como siguiente instancia a la que más disminuye la incerteza del clasificador una vez agregada al conjunto de entrenamiento. Notar que la incerteza se mide como la entropía de la clasificación.

Hasta ahora se han obtenido resultados bastante positivos, la Figura 6 muestra la exactitud del clasificador a medida que van agregándose instancias con el proceso de aprendizaje activo. La línea azul corresponde a la precisión[1] y la línea roja al recall[2]. La línea recta bajo la línea azul corresponde a la

Figura 6

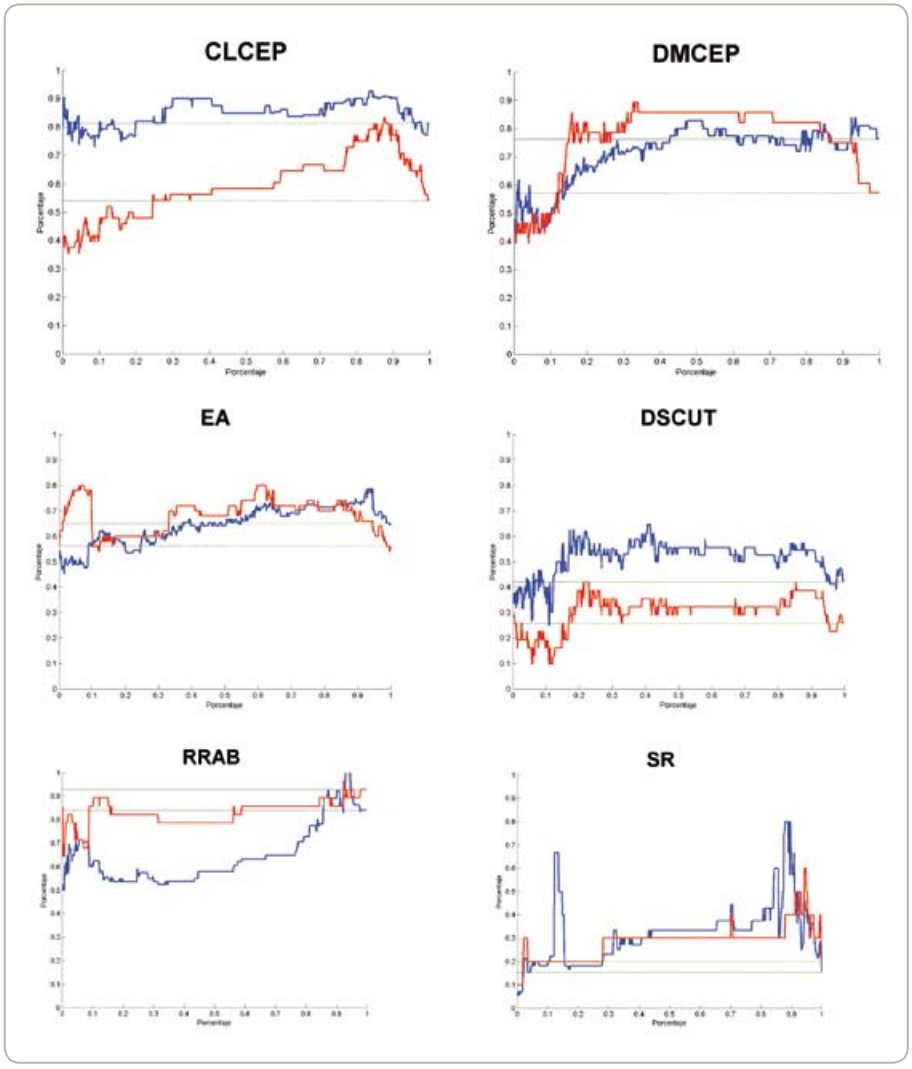


Gráfico que muestra la exactitud del clasificador a medida que van agregándose instancias con el proceso de aprendizaje activo. La línea azul corresponde a la precisión y la línea roja al recall. La línea recta bajo la línea azul corresponde a la precisión obtenida usando todo el conjunto de entrenamiento. La línea recta bajo la línea roja corresponde al recall obtenido usando todo el conjunto de entrenamiento. Cada vez que la línea roja (azul) está por sobre su línea recta implica que el recall (precisión) es mayor que el obtenido con el 100% del conjunto de entrenamiento.

1 Precisión: de los elementos que el clasificador dijo que eran de la clase en cuestión, cuántos realmente eran.
2 Recall: de los elementos de la clase que el clasificador tenía que identificar, cuántos identificó.

Susana Eyheramendy

Karim Pichara B.

Cristóbal Berger

Rodolfo Angeloni

Andrés Jordán

Márcio Catelan

precisión obtenida usando todo el conjunto de entrenamiento. La línea recta bajo la línea roja corresponde al recall obtenido usando todo el conjunto de entrenamiento. Cada vez que la línea roja (azul) está por sobre su línea recta implica que el recall (precisión) es mayor que el obtenido con el 100% del conjunto de entrenamiento. Se puede apreciar por ejemplo que en el gráfico de la clase CLCEP el clasificador logra igualar en recall y precisión los resultados obtenidos cuando se usó el 100% del conjunto de entrenamiento sólo usando un 30% de instancias, seleccionadas estratégicamente con el proceso de aprendizaje activo. Resultados similares se ven en las otras clases, excepto en la clase RRAB, donde el clasificador no puede igualar los resultados sino hasta llegar a seleccionar todas las instancias del conjunto de entrenamiento.

Los pasos siguientes de esta investigación comprenden desarrollar un modelo para automatizar la decisión de dónde detener el proceso de aprendizaje activo, es decir, en qué momento el clasificador debe decidir que ya aprendió lo suficiente y no necesita pedir la clasificación de más instancias. Para lograr este objetivo existen muchos desafíos por superar, entre ellos la inestabilidad de los resultados en algunos casos.

Existen muchas otras aristas de investigación que se irán explorando con el tiempo, este grupo científico pretende seguir creciendo y desarrollando nuevas tecnologías para el análisis de Datos Astronómicos, se espera en un futuro próximo poder contar con un número importante de estudiantes de posgrado realizando sus investigaciones en el Centro de Astro-Ingeniería de la UC, desarrollando nuevas tecnologías para la exploración eficiente de toda la información que se viene en los próximos diez años con la instalación de los nuevos observatorios en nuestro país. BITS

**Rodolfo Angeloni está financiado por el Proyecto Fondecyt #3100029.*

REFERENCIAS

Percival, D. and Andrew T. Walden. (1993) Spectral Analysis for Physical Applications. Cambridge University Press

Bishop, C. (2006) Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8.

Bloomfield, P. (1976). Fourier analysis of time series: An introduction. New York: Wiley.

Billères, M., Fontaine, G., Brassard, P., Charpinet, S., Liebert, J., Saffer, R. A., 2000, ApJ, 530, 441.

Cortes, C. and Vapnik ,V. Support vector networks. Machine Learning, 20:273–297, 1995.

Hamilton, J. (1994), Time Series Analysis, Princeton: Princeton Univ. Press, ISBN 0-691-04289-6

Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2.

Mills, Terence C. (1990) Time Series Techniques for Economists. Cambridge University Press

N. Roy and A. McCallum. Toward optimal active learning through sam- pling estimation of error reduction. In Proceedings of 18th International Conference on Machine Learning, ICML, pages 441–448, 2001.

N. Cebron and M. Berthold. Active learning for object classification: from exploration to exploitation. Data Mining and Knowledge Discovery, 18(2):283–299, 2008.

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106.

Russell, S.J. and Norvig, P. Artificial intelligence: a modern approach. Prentice Hall series in artificial intelligence. 2010

Y, Baram, R. El-Yaniv, K. Luz. Online Choice of Active Learning Algorithms. JMLR 5:255-291, 2004.

Terrell, D., Mukherjee, J.D., and Wilson, R.E. 1991. "Binary Stars: A Pictorial Atlas", Krieger Publ. Co. (Malabar, Florida).

Wachmann, A.A., Popper, D.M., and Clausen, J.V. 1986, A&A, 162, 62.