Detección de cumulos estelares en galaxias cercanas utilizando machine learning y algoritmos de redes sociales

Esp. Ing. Martin Casatti
UTN Regional Córdoba

28 de septiembre de 2022

1. Marco teórico o estado actual del tema

Las agrupaciones estelares han sido objetos reconocidos desde hace tiempo como laboratorios importantes para la investigación astrofísica. Ellas son muy útiles en varios aspectos, entre los que se pueden destacar los siguientes:

- Los cúmulos estelares contienen muestras estadísticamente significativas de estrellas de aproximadamente de la misma edad con un rango amplio de masas estelares y localizadas en un volumen relativamente pequeño del espacio, haciéndolas un conjunto ideal para el análisis de características comunes y determinación de los patrones que rigen su surgimiento [2].
- En relación con el proceso de formación estelar, los cúmulos jóvenes permiten escalrecer la forma y las escalas de tiempo en las que estos mecanismos están activos [3], así como sobre su dependendencia de los distintos ambientes interestelares de la Vía Láctea o de otras galaxias [4].

Los trabajos mencionados se han focalizado en mejorar el conocimiento de nuestra propia Galaxia (y de las Nubes de Magallanes[6]), pero actualmente hay varios factores que incrementan de forma importante tanto la cantidad de objetos a investigar cómo la metodología para hacerlo.

En la actualidad existe una gran cantidad de información de las galaxias cercanas (a varios Mpc[1]) debido, en gran parte, a que el Telescopio Espacial Hubble (HST) ha permitido obtener datos con alta resolución espacial utilizando varias cámaras de campo amplio (WFPC2; ACS) [5].

[1]Megaparsec, medida de distancia, aproximadamente 3.26 millones de años luz

Existe una enorme cantidad de datos proveniente de las varias observaciones continuas que se están realizando y que se proyectan realizar en modo "survey" (p.e. VVV, https://vvvsurvey.org/ o LSST; https://www.lsst.org/) que necesitan ser estudiados con métodos automáticos. En particular para la identificación y parametrización de nuevas agrupaciones estelares.

Los algoritmos de reconocimiento automático de patrones con el fin de encontrar agrupaciones estelares están en la actualidad teniendo una importante revisión y desarrollo, dado que su uso es vital para el análisis de los "surveys" que se están realizando. Puede encontrarse en Schmeja,(2011) [1] un "review" de los diferentes métodos y sus técnicas.

Tal como se desprende de esa publicación, estos algoritmos se basan en analizar sólo las posiciones espaciales para encontrar a los sistemas estelares por sobre-densidades contra el fondo estelar o por su equivalente relacionado con la distribución de distancias entre estrellas. Cabe hacer notar que ya se han desarrollado varios algoritmos que han sido aplicados con éxito en otros campos científicos. Entre estos se destacan algoritmos como "K-mean", "Birch", "Spectral Clustering", "Dbscan", etc.

Por otra parte, el gran auge que tienen desde hace algunos años el análisis de redes sociales nos ha brindado otro gran campo de estudios en los que se conjungan algunos de los atributos comunes que manifiesta la detección de cumulos estelares, como ser:

- En el ámbito de las redes sociales también se cuenta con una gran cantidad de datos
- Existe un conjunto de relaciones no evidentes entre los mismos y
- Un nutrido grupo de atributos analizables a fin de intentar la detección de patrones

La estructura inherente de dichas redes es la de grafos, dirigidos o no, y sobre las mismas se pueden realizar multitud de análisis sustentados por la Teoría de Grafos [7], siendo este un campo ampliamente estudiado, tanto de manera analítica como algorítmica.

Diversos estudios, tanto de la topología de dichas redes [8] como de las características que presentan sus participantes, nos brindan un fertil campo para el estudio de algoritmos de detección de patrones estructurales, muchos de ellos asistidos por técnicas de Machine Learning [9].

1.1. El estudio de la aplicación de algoritmos sobre redes sociales

En la actualidad el análisis de algoritmos y su aplicación para la determinación de las características de las redes sociales es un campo en permanente evolución.

Algoritmos como los de detección de comunidades[15], detección de anomalías[16], determinación de subredes similares, clustering dinámico[17] y predicción de enlaces más probables[18], son un ámbito en donde las técnicas de aprendizaje supervisado está encontrando cada vez más y mejores aplicaciones.

El entrenamiento de modelos específicos para la detección de este tipo de estructuras está dando lugar a cada vez más y mejores caracterizaciones de redes con una enorme cantidad de nodos y de relaciones, y abriendo el desarrollo a algoritmos más complejos y potentes.

Existen actualmente estudios comparativos de diversos algoritmos de detección de comunidades en redes [19] que presentan resultados prometedores para la aplicación de dichos algoritmos, o derivaciones de los mismos, en ámbitos diferentes, tal como es el enfoque del presente trabajo.

2. Aporte original al tema

Es la intención de este trabajo de posgrado demostrar la viabilidad de la aplicación de técnicas diseñadas para la caracterización de redes sociales, en el ámbito de la astronomía, para la detección de cumulos estelares, aprovechando de esta manera los estudios existentes en la materia pero enfocados en un nuevo ámbito de aplicación.

Se postula que la aplicación de tecnicas de machine learning para el entrenamiento de algoritmos inteligentes posibilitará que los algoritmos de detección y caracterización de comunidades en redes sociales, pueden detectar agrupaciones estelares, a partir del correspondiente cambio en los atributos descriptivos y estructurales, de acuerdo al nuevo ámbito de aplicación.

Objeto de estudio

El objeto de estudio, en particular, serán las galaxias espirales o irregulares, cercanas a la Vía Láctea, las cuales cuentan con una cantidad apreciable de estrellas azules, de gran importancia para la comunidad astronómica ya que son estrellas jóvenes en estadíos iniciales de evolución.

3. Objetivos

El presente trabajo tiene como finalidad demostrar la viabilidad de la utilización de técnicas algorítmicas de aplicación en el ámbito de redes sociales, específicamente las asociadas a comunidades de individuos, para la detección de agrupaciones estelares en galaxias cercanas.

Se analizará la viabilidad de dichas técnicas y se contrastarán los resultados obtenidos con respecto a los de otras técnicas, diseñadas específicamente para el ámbito astronómico, a fin de sacar conclusiones adecuadas al ámbito de aplicación específico.

A su vez se plantean algunos objetivos particulares a alcanzar:

- Realizar una revisión sistemática del estado del arte en cuanto a algoritmos de detección de estructuras en el ámbito astronómico y de las redes sociales.
- Determinar la viabilidad de extrapolar algoritmos de uno de los ámbitos mencionados al otro, específicamente en lo que respecta a detección de estructuras determinadas, sobre estructuras de tipo grafo.
- Establecer los atributos mínimos necesarios para el entrenamiento de un algoritmo de detección asistido por machine learning.
- Obtener un modelo de machine learning confiable para la detección de estructuras estelares, en el ámbito específico de aplicación.

4. Metodología

Para alcanzar el objetivo final mencionado, se pretende realizar las siguientes acciones:

- Se analizarán las técnicas de reconocimiento de agrupaciones estelares existentes y sus resultados actuales.
- Se analizarán las técnicas de reconocimiento de comunidades en redes sociales y sus resultados actuales.
- Se determinarán los atributos entrenables por medio de técnicas de machine learning en el ámbito de las redes sociales y extrapolarlos al ámbito astronómico.
- Se modelará y entrenará un mecanismo de machine learning con los atributos astronómicos, ya sean mediciones reales o sus equivalentes simulados.
- Se utilizará el algoritmo, una vez entrenado, para detección de comunidades sobre muestras reales a fin de analizar su eficacia y eficiencia.
- Se elaborará un procedimiento general para el entrenamiento del algoritmo de detección y la aplicación de la técnica para su utilización en diferentes ámbitos astronómicos o con diferentes muestras.

5. Resultados esperados

- Se espera, al concluir con el trabajo, contar con un modelo eficaz para la detección de agrupaciones estelares, en muestras de datos reales, con un grado de exactitud al menos comparable a los mecanismos actualmente utilizados en la comunidad astronómica para esa misma finalidad.
- Se espera demostrar que los algoritmos desarrollados para la detección de comunidades, sobre grafos de redes sociales, con las modificaciones pertinentes, pueden ser una buena alternativa a la detección de comunidades en un ámbito completamente diferente, como es el de las estrellas en galaxias cercanas.
- Se espera sentar las bases para el estudio continuo de tecnicas no desarrolladas específicamente para el ámbito astronómico, pero de posible aplicación en el mismo.
- Se espera ayudar a la comunidad astronómica con una herramienta de simple implementación y que provea resultados valiosos, como complemento a las técnicas ya existentes.

6. Antecedentes

En los últimos años se han realizado, en la unidad ejecutora, diversos estudios con respecto a la aplicación de grafos con diversos fines, entre ellos:

- “Análisis cienciométrico de la producción en investigación científica y tecnológica en la Red de Ingeniería en Informática y sistemas de información de CONFEDI”.
  Tipo de proyecto: UTN (PID UTN),
  código identificador del proyecto: 7848,
  Director: Roberto Muñoz.
  Lugar: CIDS: Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información. Facultad Regional Córdoba de la Universidad Tecnológica Nacional.
  Período 2020 : En ejecución.
- “Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico : Fase II”.
  Tipo de proyecto: UTN (PID UTN),
  código identificador del proyecto: SIUTICO0007786OO,
  Director: M. Alejandra Paz Menvielle.

Lugar: CIDS : Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información. Facultad Regional Córdoba de la Universidad Tecnológica Nacional.
Período 2020 - 2021.

- "Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico".
  Tipo de proyecto: UTN (PID UTN),
  código identificador del proyecto: SIUTNCO4812,
  Director: M. Alejandra Paz Menvielle.
  Lugar: CIDS : Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información. Facultad Regional Córdoba de la Universidad Tecnológica Nacional.
  Período 2018 - 2019.

- "Metodología para determinar la exactitud de una respuesta, escrita en forma textual, a un interrogante sobre un tema específico, aplicando herramientas informáticas".
  Tipo de proyecto: UTN (PID UTN),
  código identificador del proyecto: EIUTNCO0003592,
  Director: Mario Alberto Groppo.
  Lugar: CIDS : Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información. Facultad Regional Córdoba de la Universidad Tecnológica Nacional.
  Período 2015-2017.

El postulante ha participado en estos proyectos en calidad de docente investigador, arquitecto de las soluciones, programador, co-autor de artículos y expositor en diversos congresos y encuentros científicos.

- Text format written questions evaluation Methodology[10]
- Caso de aplicación de representación del conocimiento utilizando grafos conceptuales en un sistema de corrección automatizado de exámenes[11]
- Model and evaluation tool using graphs as knowledge base for the automated correction of exams in text format[12]
- Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico[13]
- Análisis cienciométrico de la producción en investigación científica y tecnológica[14]

7. Aportes potenciales

Contribución al avance del conocimiento científico y/o tecnológico

El proyecto está focalizado en la elaboración de un modelo de detección de patrones estelares a partir de trabajos previos desarrollados sobre comunidades de redes sociales.

La demostración de viabilidad de dichas técnicas permitirá ampliar el espectro de herramientas utilizables para la detección de cumulos estelares y propiciará el estudio de la aplicación de técnicas similares en ámbitos diversos.

Asimismo permitirá establecer la validez de ciertas técnicas de detección de patrones en grafos de cualquier tipo sobre un conjunto de datos astronómicos.

Contribución a la formación de recursos humanos

El Ing. Martin Casatti se desempeña actualmente como docente investigador en el grupo dirigido por el Ing. Roberto Muñoz, Secretario Académico de UTN Facultad Regional Córdoba, que trabaja en el marco del Centro de Investigación, Desarrollo y Transferencia de Sistemas, dirigido por el Dr. Ing. Marcelo Marciszack, director propuesto para este trabajo de posgrado.

El trabajo de la presente propuesta se desarrollará en el marco de los grupos de investigación pertenecientes a dicho Centro.

Transferencia prevista de los resultados, aplicaciones o conocimientos derivados del proyecto

La transferencia de los resultados o conocimientos del proyecto se realizará, a nivel profesional, por medio de publicaciones internacionales bajo referato, presentaciones en reuniones científicas y formación de recursos humanos.

Los resultados se compartirán y/o transferiran a instituciones relacionadas con la astronomía que estén interesadas en la aplicación de las técnicas aquí desarrolladas.

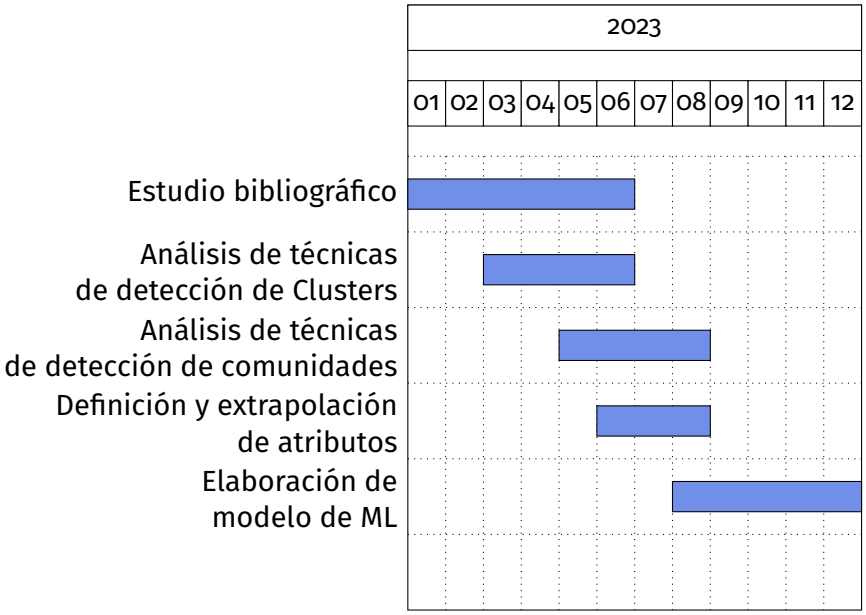
8. Director y co-director del trabajo

Para la realización del presente proyecto se postula como director y co-director, respectivamente, a los Dr. Marcelo Martín Marciszack y al Dr. Carlos Feinstein Baigorri, de los cuales se adjunta su currículum vitae detallado.

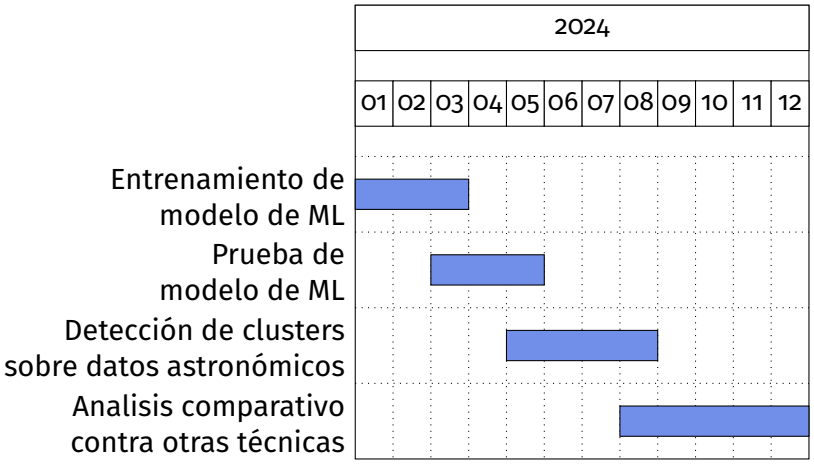
El Dr. Marcelo Marciszack se desempeña actualmente como Director del Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información (CIDS) en la Universidad Tecnológica Nacional, Facultad Regional Córdoba (Argentina), se encuentra categorizado A en la Carrera de Docente Investigador de la UTN - Orientación Ciencias de la Ingeniería y Tecnológicas, y Categorizado I en Programa de Incentivos del Ministerio de Ciencia y Tecnología.

El Dr. Carlos Feinstein es Dr. en Astronomía, por la Faculta de Ciencias Astronómicas y Geofísicas de la Universidad Nacional de La Plata (Argentina), es profesor Titular concursado en la Cátedra de Computación de la Facultad de Ciencias Astronómicas y Geofísicas, de la Universidad Nacional de La Plata, además de Investigador Independiente de CONICET desde noviembre de 2010 hasta la fecha.

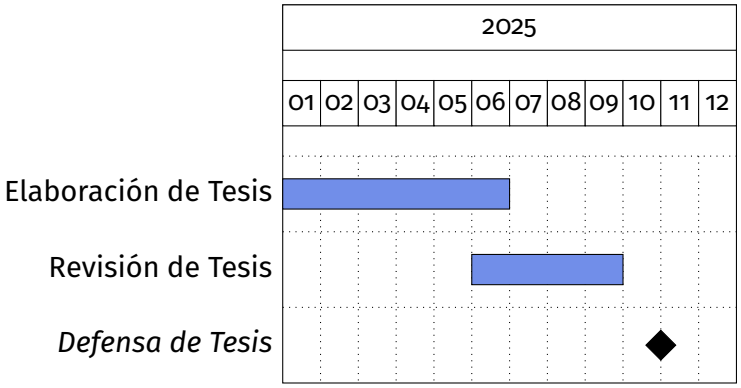
9. Plan general de trabajo



Primer año



Segundo año



Tercer año

Referencias

[1] S Schmeja. «Identifying star clusters in a field: A comparison of different algorithms». En: *Astronomische Nachrichten* 332.2 (2011), págs. 172-184.

[2] Ralf S Klessen y Andreas Burkert. «The Formation of Stellar Clusters: Gaussian Cloud Conditions. I.» En: *The Astrophysical Journal Supplement Series* 128.1 (2000), pág. 287.

[3] Hwankyung Sung, Michael S Bessell y See-Woo Lee. «UBVRI and Hα Photometry of the Young Open Cluster NGC 6231». En: *The Astronomical Journal* 115.2 (1998), pág. 734.

[4] S Michael Fall y Rupali Chandar. «Similarities in populations of star clusters». En: *The Astrophysical Journal* 752.2 (2012), pág. 96.

[5] Julianne J Dalcanton et al. «The ACS nearby galaxy survey treasury». En: *The Astrophysical Journal Supplement Series* 183.1 (2009), pág. 67.

[6] Ruben A Vázquez et al. «Spiral structure in the outer galactic disk. I. The third galactic quadrant». En: *The Astrophysical Journal* 672.2 (2008), pág. 930.

[7] Douglas Brent West et al. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River, 2001.

[8] John A Barnes y Frank Harary. «Graph theory in network analysis». En: *Social networks* 5.2 (1983), págs. 235-244.

[9] Afnan Alharbi y Khalid Alsubhi. «Botnet detection approach using graph-based machine learning». En: *IEEE Access* 9 (2021), págs. 99166-99180.

[10] María Alejandra Paz Menvielle et al. «Text format written questions evaluation Methodology». En: *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE. 2016, págs. 1-4.

[11] María Alejandra Paz Menvielle et al. «Caso de aplicación de representación del conocimiento utilizando grafos conceptuales en un sistema de corrección automatizado de exámenes». En: *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).* 2017.

[12] María Alejandra Paz Menvielle et al. «Model and evaluation tool using graphs as knowledge base for the automated correction of exams in text format». En: *2017 XLIII Latin American Computer Conference (CLEI)*. IEEE. 2017, págs. 1-10.

[13] María Alejandra Paz Menvielle et al. «Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico». En: *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).* 2018.

[14] Roberto M Muñoz et al. «Análisis cienciométrico de la producción en investigación científica y tecnológica». En: *XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz).* 2020.

[15] Cuijuan Wang et al. «Review on community detection algorithms in social networks». En: *2015 IEEE international conference on progress in informatics and computing (PIC)*. IEEE. 2015, págs. 551-555.

[16] Ravneet Kaur y Sarbjeet Singh. «A survey of data mining and social network analysis based anomaly detection techniques». En: *Egyptian informatics journal* 17.2 (2016), págs. 199-216.

[17] S Boccaletti et al. «Detecting complex network modularity by dynamical clustering». En: *Physical Review E* 75.4 (2007), pág. 045102.

[18] Ajay Kumar Singh Kushwah y Amit Kumar Manjhvar. «A review on link prediction in social network». En: *International Journal of Grid and Distributed Computing* 9.2 (2016), págs. 43-50.

[19] Andrea Lancichinetti y Santo Fortunato. «Community detection algorithms: A comparative analysis». En: *Phys. Rev. E* 80 (5 nov. de 2009), pág. 056117. DOI: 10.1103/PhysRevE.80.056117. URL: https://link.aps.org/doi/10.1103/PhysRevE.80.056117.