Detección de cúmulos estelares en galaxias cercanas utilizando técnicas de Machine Learning y algoritmos de aplicación en redes sociales

Universidad Tecnológica Nacional
Facultad Regional Córdoba

Plan de tesis para optar al título de Dr. en Ingeniería,
mención Sistemas de Información

Tesista: Esp. Ing. Martin Casatti
Director: Dr. Marcelo Marciszack
CoDirector: Dr. Carlos Feinstein

1. Introducción

Detección e identificación de cúmulos estelares

Las agrupaciones estelares, también denominados cúmulos o clusters, han sido objetos reconocidos desde hace tiempo como laboratorios importantes para la investigación astrofísica, siendo muy útiles en varios aspectos, entre los que se pueden destacar los siguientes:

- Contienen muestras estadísticamente significativas de estrellas de aproximadamente la misma edad, con composiciones químicas similares, un amplio rango 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 [1].

- En relación con el proceso de formación estelar, los cúmulos jóvenes permiten esclarecer la forma y las escalas de tiempo en las que estos mecanismos están activos, así como también permiten analizar su dependendencia de los distintos ambientes interestelares de la Vía Láctea o de otras galaxias [2].

Los trabajos mencionados se han focalizado en mejorar el conocimiento de nuestra propia Galaxia (y de las Nubes de Magallanes[3]), 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) [4].

Se cuenta con una enorme cantidad de datos proveniente de las varias observaciones continuas que se están realizando y que se proyectan realizar en modo "survey"[2] (p.e. VVV[3] o LSST[4]) que necesitan ser estudiados con métodos automáticos.

En este ámbito, los algoritmos de reconocimiento automático de patrones, están teniendo una importante revisión y desarrollo tal como se puede apreciar en el análisis comparativo de Schmeja (2011)[5].

En otros ámbitos científicos se han aplicado con éxito diversos algoritmos de clustering, como por ejemplo "K-mean", "Birch", "Spectral Clustering", "Dbscan", etc.[6]

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

[2] Técnica que consiste en realizar un mapeo sistemático de una porción determinada de la esfera celeste sin concentrarse de manera puntual en ningún objeto.

[3] https://vvvsurvey.org/

[4] https://www.lsst.org/

El estudio de la estructura de redes sociales

El auge que tiene desde hace algunos años el análisis de redes sociales nos ha brindado otro amplio campo de estudios en el que se pueden apreciar algunos de los atributos que son comunes al problema de la detección de cúmulos estelares, como por ejemplo:

- 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 guiar la detección de patrones.

La estructura inherente de dichas redes es la de un grafo, sobre el que se puede realizar multitud de análisis sustentados por la Teoría de Grafos [7].

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

Algoritmos como los de "detección de comunidades" [10], "detección de anomalías" [11], "determinación de subredes similares", "clustering dinámico" [12] y "predicción de enlaces más probables" [13], son un ámbito en donde las técnicas de aprendizaje supervisado está encontrando cada vez más aplicaciones.

Existen actualmente estudios comparativos de diversos algoritmos de detección de comunidades en redes [14] que presentan resultados prometedores para su aplicación, o las de sus derivados, en ámbitos diferentes, tal como es el enfoque del presente trabajo.

2. Justificación

La puesta en funcionamiento de instrumentos de observación astronómica cada vez más potentes, durante los últimos 50 años, ha dado lugar a un crecimiento exponencial de la cantidad de objetos detectados, los que requieren análisis y estudio.

Sin ir demasiado lejos, el recientemente lanzado telescopio James Webb produce casi 60 Gigabytes de información al día, la cual no puede ser almacenada de manera local y debe ser transmitida de inmediato el centro de control de misión[15], mientras que el proyecto "Legacy Survey of Space and Time" (Rubin/LSST), basado en el observatorio Vera C. Rubin[5], en Chile, se estima que producirá 20 TB (terabytes) de información cada noche, durante una vida útil de al menos 10 años [16].

Estos volúmenes de datos hacen que sea imprescindible la utilización de mecanismos automáticos para su análisis.

[5] https://rubinobs.org/

Es en este sentido que creemos que los resultados del presente trabajo pueden aportar al avance de dichas técnicas y colaborar, en última instancia, en el avance científico y tecnológico.

3. Hipótesis de trabajo

Es la intención de esta tesis doctoral de posgrado demostrar la viabilidad de la aplicación de técnicas inicialmente diseñadas para la caracterización de redes sociales, en el ámbito de la astronomía, para la detección de cúmulos 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 técnicas de machine learning para el entrenamiento de algoritmos inteligentes posibilitará que los algoritmos de detección y caracterización de comunidades en redes sociales, puedan detectar agrupaciones estelares, a partir del correspondiente cambio en los atributos descriptivos y estructurales, de acuerdo al nuevo ámbito de aplicación.*

4. Objetivos

Objetivo principal

1. El presente trabajo tiene como finalidad demostrar la viabilidad de la utilización de técnicas algorítmicas de aplicación en el ámbito de redes sociales para la detección de agrupaciones estelares en galaxias cercanas.

Objetivos secundarios

Se plantean asimismo los siguientes objetivos particulares a alcanzar:

1.1 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.

1.2 Determinar la viabilidad de extrapolar algoritmos utilizados en el ámbito de las redes sociales, para su aplicación en el ámbito astronómico, específicamente en lo que respecta a detección de estructuras determinadas, sobre estructuras de tipo grafo.

1.3 Establecer los atributos mínimos necesarios para el entrenamiento de un algoritmo de detección asistido por machine learning.

1.4 Obtener un modelo de machine learning confiable para la detección de estructuras estelares, en el ámbito específico de aplicación.

5. Metodología

Para alcanzar los objetivos de la presente tesis, tanto a nivel general general como los objetivos particulares, se realizarán las siguientes actividades:

A1 Para cumplimentar con el objetivo **1.1** se analizarán las técnicas de reconocimiento de agrupaciones estelares existentes, realizando una revisión sistemática de literatura, para determinar la efectividad de detección de cada una de ellas a fin de obtener una línea base.

A2 Para cumplimentar con el objetivo **1.1** se identificarán las principales técnicas de reconocimiento de comunidades en redes sociales, por medio de una revisión sistemática de literatura, a fines de establecer qué algoritmos son aplicables en el dominio astronómico de acuerdo a sus características.

A3 Para alcanzar el objetivo **1.2** se construirá un entorno de pruebas, con un set de datos acotado y conocido, a partir del repositorio GAIA[6] y similares (ESO[7], NASA[8], etc.), sobre el cual se realizará la aplicación de un algoritmo de detección de comunidades originalmente diseñado para redes sociales, y se analizará el resultado obtenido de su aplicación sobre un dominio astronómico, a fin de validar conceptualmente la propuesta de trabajo de esta tesis.

A4 Mediante el análisis de sets de datos de redes sociales, descargados de repositorios como el Stanford Large Network Dataset Collection[9] o el Network Data Repository[10], se determinarán los atributos entrenables por medio de técnicas de machine learning y los mismos se extrapolarán a sets de datos astronómicos para cumplimentar con los objetivos **1.2** y **1.3**.

A5 Para cumplir con el objetivo **1.3** se modelará y entrenará un mecanismo de machine learning con los atributos astronómicos, ya sean mediciones reales o sus equivalentes simulados, utilizando las librerías más populares y probadas en la actualidad, como TensorFlow, PyTorch, SciKit-Learn o Keras, para determinar la eficacia en la detección de clusters.

A6 Para cumplimentar con los objetivos **1.3** y **1.4** se utilizará el algoritmo, una vez entrenado, para detección de comunidades sobre muestras reales, como los datos obtenidos del repositorio GAIA mencionado anteriormente, analizando la cantidad de atributos reconocidos, la cantidad de agrupaciones, el tiempo de reconocimiento y otras características escenciales, a fin de determinar si la eficacia se mantiene sobre muestras de datos reales y medir su eficiencia.

[6] https://gea.esac.esa.int/archive/
[7] http://archive.eso.org/cms.html
[8] https://nssdc.gsfc.nasa.gov/astro/
[9] https://snap.stanford.edu/data/
[10] https://networkrepository.com/soc.php

A7 Para alcanzar el objetivo **1.4** 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, el cual se plasmará en un documento con instrucciones detalladas para el preprocesamiento de los atributos, la selección de los algortimos a utilizar, el entrenamiento de la red y indicaciones para el análisis de los resultados obtenidos de su aplicación.

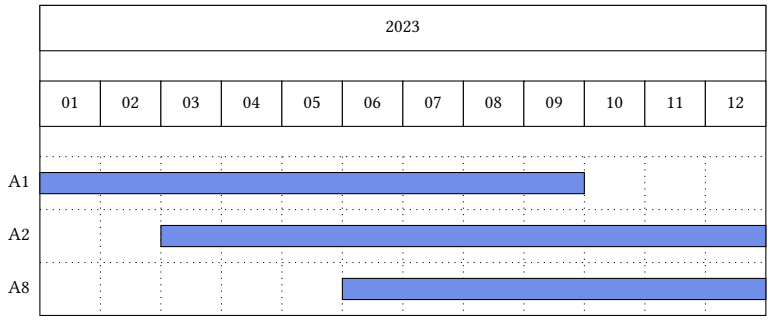
A8 Para alcanzar el objetivo general **1** se publicará de manera regular los avances y resultados obtenidos, a fin de validar los mismos con la comunidad científica.

6. Resultados esperados

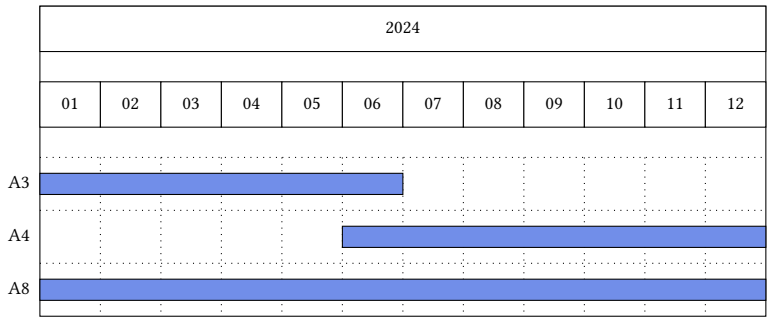
- Se espera, al concluir con el trabajo, contar con un modelo eficaz para la detección de agrupaciones estelares, en muestras de datos reales, con un grado de exactitud al menos comparable a los mecanismos actualmente utilizados en la comunidad astronómica para esa misma finalidad.
- Se espera demostrar 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 técnicas 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.

7. Cronograma y plan general de trabajo

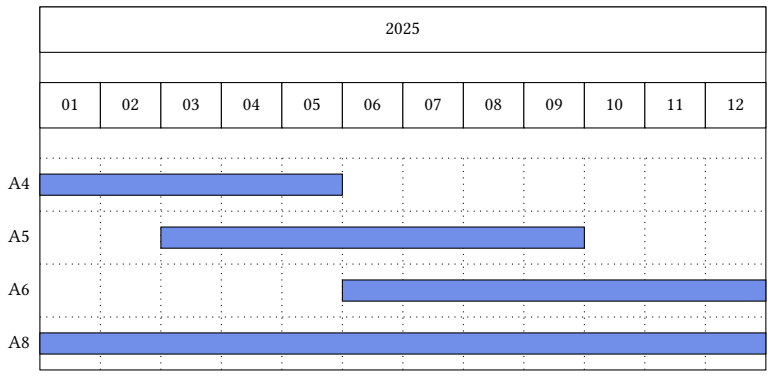
En la presente sección se presentará la calendarización de actividades, cuyo detalle se encuentra en la sección Metodología.



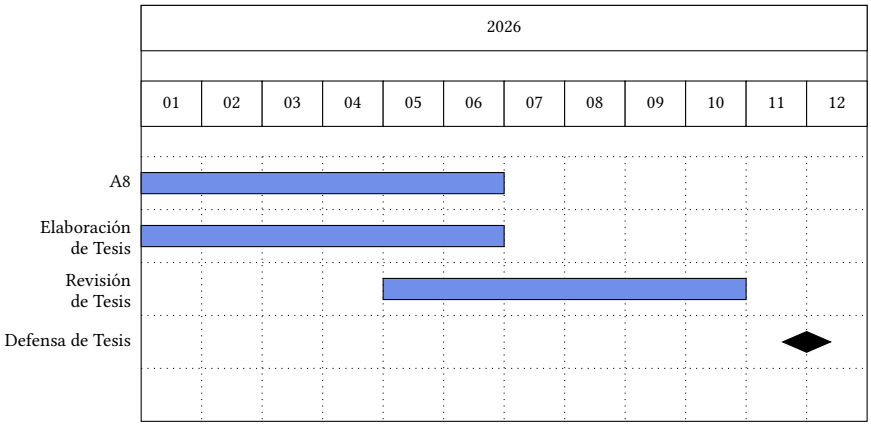
Primer año



Segundo año



Tercer año



Cuarto año

8. Antecedentes del tesista

En los últimos años se han realizado, en la unidad ejecutora (CIDS: Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información. Facultad Regional Córdoba de la Universidad Tecnológica Nacional), 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".*
  Director: Roberto Muñoz.
  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".*
  Director: M. Alejandra Paz Menvielle.
  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".*
  Director: M. Alejandra Paz Menvielle.
  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".*
  Director: Mario Alberto Groppo.
  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, algunos de los cuales se mencionan a continuación.

- Text format written questions evaluation Methodology[17]
- Caso de aplicación de representación del conocimiento utilizando grafos conceptuales en un sistema de corrección automatizado de exámenes[18]
- Model and evaluation tool using graphs as knowledge base for the automated correction of exams in text format[19]
- Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico[20]
- Análisis cienciométrico de la producción en investigación científica y tecnológica[21]

9. Aportes potenciales del trabajo

Contribución al avance del conocimiento científico y/o tecnológico

La demostración de viabilidad de dichas técnicas permitirá ampliar el espectro de herramientas utilizables para la detección de cúmulos 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, 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 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.

10. 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, y está incluído en la Categoría II del Programa de Incentivos del Ministerio de Ciencia y Tecnología.

11. Infraestructura y equipamiento

Las tareas se desarrollarán en las instalaciones del Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información, en la Universidad Tecnológica Nacional, Facultad Regional Córdoba.

Dicha locación cuenta con los requerimientos necesarios para el desarrollo del trabajo en cuanto a equipamiento informático, conectividad, acceso a bases de datos, información de referencia, etc.

Referencias

1. Klessen, R. S. y Burkert, A. The Formation of Stellar Clusters: Gaussian Cloud Conditions. I. *The Astrophysical Journal Supplement Series* **128,** 287 (2000).
2. Fall, S. M. y Chandar, R. Similarities in populations of star clusters. *The Astrophysical Journal* **752,** 96 (2012).
3. Vázquez, R. A. *et al.* Spiral structure in the outer galactic disk. I. The third galactic quadrant. *The Astrophysical Journal* **672,** 930 (2008).
4. Dalcanton, J. J. *et al.* The ACS nearby galaxy survey treasury. *The Astrophysical Journal Supplement Series* **183,** 67 (2009).
5. Schmeja, S. Identifying star clusters in a field: A comparison of different algorithms. *Astronomische Nachrichten* **332,** 172-184 (2011).
6. Rodriguez, M. Z. *et al.* Clustering algorithms: A comparative approach. *PloS one* **14,** e0210236 (2019).

7. West, D. B. *et al. Introduction to graph theory* (Prentice hall Upper Saddle River, 2001).
8. Barnes, J. A. y Harary, F. Graph theory in network analysis. *Social networks* **5,** 235-244 (1983).
9. Alharbi, A. y Alsubhi, K. Botnet detection approach using graph-based machine learning. *IEEE Access* **9,** 99166-99180 (2021).
10. Wang, C., Tang, W., Sun, B., Fang, J. y Wang, Y. *Review on community detection algorithms in social networks* en *2015 IEEE international conference on progress in informatics and computing (PIC)* (2015), 551-555.
11. Kaur, R. y Singh, S. A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian informatics journal* **17,** 199-216 (2016).
12. Boccaletti, S., Ivanchenko, M., Latora, V., Pluchino, A. y Rapisarda, A. Detecting complex network modularity by dynamical clustering. *Physical Review E* **75,** 045102 (2007).
13. Kushwah, A. K. S. y Manjhvar, A. K. A review on link prediction in social network. *International Journal of Grid and Distributed Computing* **9,** 43-50 (2016).
14. Lancichinetti, A. y Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80,** 056117. https://link.aps.org/doi/10.1103/PhysRevE.80.056117 (5 nov. de 2009).
15. *All Quick Facts* [Online; accessed 4. Feb. 2023]. https://webbtelescope.org/quick-facts/all-quick-facts.
16. Telescope, L. S. S. Key Numbers. *Rubin Observatory.* https://www.lsst.org/scientists/keynumbers (jul. de 2021).
17. Menvielle, M. A. P., Groppo, M. A., Marciszack, M. M. y Casatti, M. *Text format written questions evaluation Methodology* en *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)* (2016), 1-4.
18. Paz Menvielle, M. A. *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).
19. Menvielle, M. A. P. *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)* (2017), 1-10.
20. Paz Menvielle, M. A., Corso, C. L., Ligorria, K., Guzmán, A. y Casatti, M. *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).

21. Muñoz, R. M. *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).