

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

Resumen

El presente trabajo expone la hipótesis de trabajo y las actividades en desarrollo de una tesis de doctorado que busca demostrar la viabilidad de la utilización de técnicas de reconocimiento de comunidades en redes sociales pero aplicadas a la detección de clusters estelares en galaxias cercanas. El trabajo expone las condiciones actuales, que dan surgimiento a la necesidad del uso de técnicas de detección automatizada, describe los ámbitos bajo estudio y plantea la hipótesis de trabajo así como las tareas a desarrollar para la consecución del objetivo general de la tesis.

El presente trabajo se enmarca en el desarrollo de una tesis doctoral, llevada adelante en forma conjunta por el Centro de Investigación, Desarrollo y Transferencia de Sistemas, de la Universidad Tecnológica Nacional, Facultad Regional Córdoba, con el Instituto de Astrofísica de la Universidad Nacional de La Plata y cuyo objetivo principal es lograr la detección de cúmulos estelares en galaxias cercanas utilizando algoritmos de redes sociales y técnicas de Machine Learning.

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 (Klessen y Burkert, 2000).
- 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 dependencia de los distintos ambientes interestelares de la Vía Láctea o de otras galaxias (Fall y Chandar, 2012).

En la actualidad existe una gran cantidad de información de las galaxias cercanas (a varios Mpc¹) 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) (Dalcanton et al., 2009).

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. VVV³ o LSST⁴) 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)(Schmeja, 2011).

En otros ámbitos científicos se han aplicado con éxito diversos algoritmos de clustering, como por ejemplo “K-mean”, “Birch”, “Spectral Clustering”, “Dbscan”, etc.(Rodríguez et al., 2019)

Por otra parte, 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 (West et al., 2001).

Diversos estudios, tanto de la topología de dichas redes (Barnes y Harary, 1983) como de las características que presentan sus participantes, brindan un fértil campo para el estudio de algoritmos de detección de patrones estructurales, muchos de ellos asistidos por técnicas de Machine Learning (Alharbi y Alsubhi, 2021).

Algoritmos como los de “detección de comunidades” (Wang et al., 2015), “detección de anomalías” (Kaur y Singh, 2016), “determinación de subredes similares”, “clustering dinámico” (Boccaletti et al., 2007) y “predicción de enlaces más probables” (Kushwah y Manjhvar, 2016), 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 (Lancichinetti y Fortunato, 2009) que presentan resultados prometedores para su aplicación, o las de sus derivados, en ámbitos diferentes.

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

²Técnica 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/>

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 al centro de control de misión (*All Quick Facts* 2023), mientras que el proyecto “Legacy Survey of Space and Time” (Rubin/LSST), basado en el observatorio Vera C. Rubin⁵, en Chile, se estima que producirá 20 TB (terabytes) de información cada noche, durante una vida útil de al menos 10 años (Telescope, 2021).

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

La tesis 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.

Para ellos se analizarán tanto las técnicas actuales de reconocimiento de agrupaciones estelares, como técnicas actuales para reconocimiento de comunidades en redes sociales. Se prevé construir un entorno de pruebas, basado en datos del repositorio GAIA⁶ o similar (ESO⁷, NASA⁸, 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.

Por otra parte se realizará un análisis de sets de datos de redes sociales, descargados de repositorios como el Stanford Large Network Dataset Collection⁹ o el Network Data Repository¹⁰, en base al cual se determinarán los atributos entrenables por medio de técnicas de machine learning y los mismos se extrapolarán a sets de datos astronómicos.

Por último 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. Este modelo se pondrá a prueba, mediante su aplicación a los datos de

⁵<https://rubinobs.org/>

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>

GAIA previamente mencionados, y se analizará la eficacia y precisión de la detección de cúmulos estelares que se hubiera alcanzado.

Referencias

- Alharbi, Afnan y Khalid Alsubhi (2021). «Botnet detection approach using graph-based machine learning». En: *IEEE Access* 9, págs. 99166-99180.
- All Quick Facts* (feb. de 2023). [Online; accessed 4. Feb. 2023]. URL: <https://webbtelescope.org/quick-facts/all-quick-facts>.
- Barnes, John A y Frank Harary (1983). «Graph theory in network analysis». En: *Social networks* 5.2, págs. 235-244.
- Boccaletti, S et al. (2007). «Detecting complex network modularity by dynamical clustering». En: *Physical Review E* 75.4, pág. 045102.
- Dalcanton, Julianne J et al. (2009). «The ACS nearby galaxy survey treasury». En: *The Astrophysical Journal Supplement Series* 183.1, pág. 67.
- Fall, S Michael y Rupali Chandar (2012). «Similarities in populations of star clusters». En: *The Astrophysical Journal* 752.2, pág. 96.
- Kaur, Ravneet y Sarbjeet Singh (2016). «A survey of data mining and social network analysis based anomaly detection techniques». En: *Egyptian informatics journal* 17.2, págs. 199-216.
- Klessen, Ralf S y Andreas Burkert (2000). «The Formation of Stellar Clusters: Gaussian Cloud Conditions. I.» En: *The Astrophysical Journal Supplement Series* 128.1, pág. 287.
- Kushwah, Ajay Kumar Singh y Amit Kumar Manjhvar (2016). «A review on link prediction in social network». En: *International Journal of Grid and Distributed Computing* 9.2, págs. 43-50.
- Lancichinetti, Andrea y Santo Fortunato (nov. de 2009). «Community detection algorithms: A comparative analysis». En: *Phys. Rev. E* 80 (5), pág. 056117. DOI: 10.1103/PhysRevE.80.056117. URL: <https://link.aps.org/doi/10.1103/PhysRevE.80.056117>.
- Rodriguez, Mayra Z et al. (2019). «Clustering algorithms: A comparative approach». En: *PloS one* 14.1, e0210236.
- Schmeja, S (2011). «Identifying star clusters in a field: A comparison of different algorithms». En: *Astronomische Nachrichten* 332.2, págs. 172-184.
- Telescope, Large Synoptic Survey (jul. de 2021). «Key Numbers». En: *Rubin Observatory*. URL: <https://www.lsst.org/scientists/keynumbers>.
- Wang, Cuijuan et al. (2015). «Review on community detection algorithms in social networks». En: *2015 IEEE international conference on progress in informatics and computing (PIC)*. IEEE, págs. 551-555.
- West, Douglas Brent et al. (2001). *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River.