CARACTERIZACIÓN DE REDES ESTELARES CON HERRAMIENTAS DE ANALÍTICA DE DATOS Y MODELADO DE GRAFOS

Martín Gustavo Casatti[1], Marcelo Martín Marciszack[1] y Carlos Feinstein Baigorri[2]
[1]Universidad Tecnológica Nacional – Facultad Regional Córdoba
Maestro Marcelo Lopez esq. Cruz Roja Argentina, Córdoba, Argentina
[2]Universidad Nacional de La Plata - Instituto de Astrofísica de La Plata
Avenida Centenario (Paseo del Bosque) S/N, La Plata, Argentina

RESUMEN

El presente trabajo expone los resultados obtenidos a partir del pre procesamiento, estructuración y posterior análisis de datos astronómicos a los fines de generar una red de datos asociados a observaciones de estrellas en galaxias cercanas, más específicamente en las Nubes de Magallanes a fin de caracterizar dicha estructura y comprobar sus similitudes y diferencias con respecto a una estructura conocida y ampliamente analizada denominada "Redes de Mundo Pequeño", estructura que reviste una especial importancia a la hora de seleccionar algoritmos de detección de comunidades, los cuales funcionan de manera particularmente eficiente cuando el grafo subyacente presenta una estructura de "Mundo Pequeño". Se presentan todos los pasos previos de adecuación y estandarización de datos, los filtros aplicados para eliminar información redundante o innecesaria, la determinación de atributos y datos relacionables, la construcción del grafo subyacente y el posterior análisis del mismo. Por último, se exponen las conclusiones obtenidas y los próximos pasos a seguir en la investigación.

PALABRAS CLAVE

Clúster, Patrón, Estructura de Red, Mundo Pequeño, Grafo, Analítica

1. INTRODUCCIÓN

1.1 El Estudio de Agrupaciones 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 ellos:

- Contienen muestras estadísticamente significativas de estrellas de aproximadamente la misma edad, 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).
- 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 (Fall y Chandar 2012).

Los trabajos mencionados se han focalizado en mejorar el conocimiento de nuestra propia Galaxia (y de las Nubes de Magallanes (Vázquez et al. 2008), pero actualmente hay varios factores que incrementan de forma importante la cantidad de objetos a investigar.

Se cuenta con una enorme cantidad de datos proveniente de las varias observaciones continuas que se están realizando en modo 'survey' (p.e. VVV o LSST) 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 (Schmeja 2011). Estos algoritmos se basan en analizar las posiciones espaciales para encontrar los clusters estelares por sobre densidades contra el fondo estelar o por su equivalente relacionado con la distribución de distancias entre estrellas.

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).

1.2 Comunidades y Redes Sociales

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.
- 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, nos 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).

Actualmente, el análisis de algoritmos y su aplicación para determinar las características de las redes sociales es un campo en permanente evolución. 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 están encontrando cada vez más aplicaciones.

Uno de los requisitos para la aplicación de varios de los algoritmos mencionados es que la red cumpla con los requisitos de ser una ‘‘red de mundo pequeño’’, característica que se analizará más adelante en este trabajo.

1.3 Trabajos Previos

El trabajo ‘‘Bases de datos de grafos como soporte para la detección de estrellas jóvenes en cúmulos estelares cercanos’’ publicado en la 7º Conferencia de Big Data y Cloud Computing (Cloud Computing and Big Data s.f.), sentó las bases iniciales para la línea de investigación del presente trabajo. El mismo utiliza algunos conceptos analizados sobre otro ámbito de aplicación, como son las bases de datos cienciométricas y las características que debe reunir un sistema de almacenamiento de información para permitir la detección de patrones de manera eficiente. Dicho enfoque se expone en los trabajos ‘‘Criterios para el diseño de una base de datos cienciométrica’’ (Muñoz et al. s.f.) y ‘‘Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico’’ (Paz Menvielle et al. 2018).

1.4 Motivación del Presente Trabajo

Los algoritmos de agrupamiento actualmente utilizados en la mayoría de los observatorios virtuales utilizan un enfoque basado en proximidad espacial. Es decir, utilizan las posiciones relativas de las estrellas, comparadas unas con otras, para determinar posibles relaciones de pertenencia a un determinado ‘‘clúster’’.

Este enfoque, si bien efectivo, deja de lado otros atributos que caracterizan a las estrellas de los agrupamientos y que podrían ser tomados en cuenta, como la luminosidad, composición química, características de movimiento, etc. (Karttunen et al. 2007; Lang y Lang 2013). El modelado de las estructuras estelares en forma de grafo, donde los atributos se expresan como relaciones entre nodos, busca resolver la representación de los atributos mencionados y la utilización de estos como criterio de agrupamiento.

Por otra parte, los algoritmos de detección de comunidades en redes sociales utilizan desde sus inicios atributos muy variados para la construcción de modelos y la detección de agrupamientos, destacando que muchos de estos algoritmos se basan precisamente en esta información adicional para su funcionamiento (Kumar, Chawla y Rana 2018). Existen actualmente multitud de algoritmos de probada efectividad sobre redes sociales (Wang et al. 2015), los cuales pueden ser explotados en un entorno astronómico para complementar las herramientas existentes y ampliar el arsenal de técnicas a disposición de los astrónomos.

La caracterización de redes estelares en forma de grafo es el primer paso en la aplicación de dichas técnicas y es la motivación principal para este trabajo.

1.5 Estructura del Trabajo

El presente trabajo cuenta con una introducción, en la sección 1, en la cual se mencionan los estudios principales en los cuales se basó y la motivación subyacente, a continuación, en la sección 2 se menciona todo el proceso de adquisición de datos, preprocesamiento y construcción y análisis del grafo estelar resultante, en la sección 3 se presentan los resultados obtenidos y, finalmente, en la sección 4 se exponen las conclusiones y se mencionan las posibles líneas futuras de investigación.

2. DESARROLLO

Las ‘‘redes de mundo pequeño’’ son un tipo especial de grafo, dirigido o no, cuyas características más importantes se pueden resumir en:

1. Tienen un coeficiente de agrupamiento alto (clustering coefficient).
2. Tienen una longitud de camino promedio corto (average path length).

Esto coincide con observaciones realizadas sobre diversos tipos de redes sociales, las cuales tienen comunidades muy conectadas separadas entre sí por enlaces débiles. En las siguientes secciones se analizarán los pasos para determinar si una red estelar puede considerarse una ‘‘de mundo pequeño’’ con lo que implica para analizar patrones y comunidades.

2.1 Obtención y Preprocesamiento de Datos con TOPCAT

Para obtener los datos iniciales se consultó por nombre sobre el catálogo GAIA versión 3, para la Pequeña Nube de Magallanes (SMC o NGC292), que estudia este trabajo. Accediendo a las interfaces de consulta del catálogo se realizó una búsqueda por nombre la cual devolvió 3,024,418 registros, los cuales deberían ser preprocesados para reducir el set de datos y eliminar información redundante e innecesaria, tarea que se realizó utilizando la herramienta TOPCAT de gestión de tablas astronómicas.

Como el análisis se realizaría sobre los datos del movimiento propio de las estrellas (proper motion o PM, en adelante) se debió diferenciar las estrellas más lejanas, con un movimiento propio imperceptible, de aquellas más cercanas, las que cuentan con un movimiento claramente detectable con los instrumentos apropiados, para lo cual se calcularon las métricas estadísticas sobre las columnas de datos asociadas al movimiento propio en sus dos componentes, ascensión recta (rect ascention, RA) y declinación (declination, DEC).

Después se realizó un filtrado de datos descartando aquellas estrellas cuyo movimiento fuera inferior a la media de los movimientos del set de datos en general y aquellas cuyo error de lectura superase al error de lectura promedio del set de datos, en valor absoluto para evitar compensaciones. Este proceso redujo el set de datos a 180,900 datos, los cuales se cargaron en una notebook Jupyter para seguir su procesamiento por medio de Python y AstroPy.

2.2 Procesamiento Utilizando AstroPy

Luego de importados los datos a una Jupyter Notebook se volvieron a calcular los indicadores estadísticos descriptivos sobre la nueva muestra, entre ellos la media del movimiento propio, dato bajo estudio y, una vez obtenido dicho valor de todas las estrellas del set de datos se tomó como referencia un rango de ±20 % por

encima y por debajo de dicho promedio y se individualizaron las estrellas que cumplieran con este requisito, para su análisis detallado.

Ya habiendo obtenido las estrellas con movimientos similares dentro del set de datos, se procedió a detectar cuáles de ellas estaban próximas entre sí. El procedimiento utilizado fue el siguiente:

1. Recorrer todas las estrellas del set de datos filtrado.
2. Para cada una de ellas, obtener un círculo con centro en la estrella y radio equivalente a 5 minutos de arco y realizar el siguiente proceso:
    a. Buscar las estrellas que se encuentran dentro del rango.
    b. Agregarlas a una lista de estrellas "vecinas".
    c. Asociar la lista de "vecinas" a la estrella sobre la cual se está realizando el análisis.
    d. Proseguir con la siguiente estrella de la lista.

Una vez completado el análisis, se cuenta con una estructura de datos similar a la de la Tabla 1:

Tabla 1. Estructura de datos

Identificador	Vecinas
6377284298571599744	[6379356298170174336]
6377285849055413888	[6379353613814721280, 6379353721189807488]

Estos datos se utilizaron después para construir el grafo para analizar la estructura y atributos de la red. Se determinó que aquellas estrellas que no cuentan con vecinas no aportan valor al análisis de la red, por lo que los elementos asociados a estrellas aisladas también se eliminaron del set de datos.

Como paso final, se exportaron los datos de la notebook Jupyter en un formato adecuado para ser importado en el software de análisis de grafos, que, en nuestro caso, se llevó a cabo utilizando Gephi.

2.3 Importación y Análisis en Gephi

Gephi es un software de análisis de redes, de código abierto y gratuito, ampliamente utilizado por su potencia y facilidad de uso.

La herramienta prevé multitud de análisis y métricas, incluye un entorno de visualización interactiva y permite la generación de grafos con características completamente definidas por el usuario. Fue de especial importancia para este trabajo la posibilidad de importar información de fuentes externas. En el presente trabajo se utilizó la versión 0.10, actualizada a noviembre de 2023.

Los archivos exportados desde Jupyter tienen la estructura esperada por Gephi para su importación, la cual se presenta en la Tabla 2.

La estructura de dichos archivos unifica nodos y aristas en una única estructura de dos columnas, en la primera de las cuales se encuentra la fuente (nodo origen) y en la segunda se encuentra el destino (nodo destino). La herramienta analiza dicha información y en caso de que alguno de los nodos no existe se crea antes de establecer la relación indicada.

En caso de que un nodo fuente, o destino, ya exista, simplemente se crea la arista que los relaciona.

Tabla 2. Estructura de importación de Gephi

source	target
6377284298571599744	6379356298170174336
6377285849055413888	6379353613814721280
6377285849055413888	6379353721189807488

Una vez generados los archivos desde Jupyter, los mismos se importan a Gephi, momento en el cual el software crea todos los nodos necesarios y establece las relaciones (aristas) entre los nodos "source" y "target".

Tabla 3. Características del grafo bajo estudio

Atributo	Grafo Aleatoria (referencia)	Grafo de Movimiento Propio (bajo análisis)
Diámetro	16	9
Radio	1	1
Cantidad de nodos	2766	2936
Cantidad de aristas	4221	4076

Para determinar las características propias del grafo estelar, se generó un grafo comparativo utilizando información aleatoria y generando aproximadamente los mismos nodos y relaciones que el grafo analizado, para comparar sus estructuras y características, entre ellas:

- Diámetro: La mayor distancia entre dos nodos de un grafo.
- Cantidad de nodos: Cantidad total de nodos de la estructura. Empíricamente se considera que dos grafos con hasta un 10% de similitud en la cantidad de nodos son lo suficientemente similares como para que el análisis fuera válido (en este caso, 6%).
- Cantidad de aristas: Cantidad total de aristas de la estructura. Empíricamente se considera que dos grafos con hasta un 10% de similitud en la cantidad de aristasson lo suficientemente similares como para que el análisis fuera válido (en este caso, 3%).

Los atributos de ambos grafos se resumen en la Tabla 3.

La visualización de ambos grafos se puede apreciar en las Figuras 1a y 1b, en donde se pueden distinguir algunos agrupamientos que serán analizados en mayor detalle en la sección 3.

3. RESULTADOS

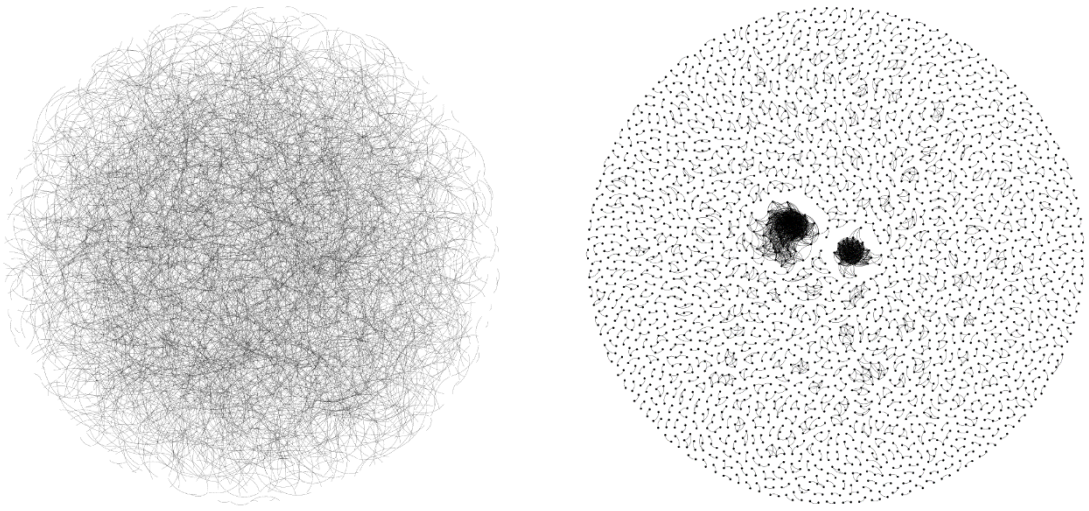
El criterio más aceptado para determinar si una red dada puede considerarse una “red de mundo pequeño” radica en la comparación de dos valores típicos de las métricas de grafos (Newman, Barabási y Watts 2011). La red tiene:

- Un coeficiente de agrupamiento superior a un grafo aleatorio de igual tamaño.
- Una longitud de camino promedio menor que en un grafo aleatorio de igual tamaño.

Dicho de otra manera, la red está más agrupada y presenta comunidades compactas, separadas unas de otras por caminos de longitud elevada. En los dos grafos presentados anteriormente se puede apreciar este fenómeno, corroborado por las métricas que se muestran en la Tabla 4.

Tabla 4. Métricas comparativas del grafo aleatorio vs. Red Estelar

	Grafo aleatorio	Red estelar bajo estudio
Coeficiente de agrupamiento	0.001	0.606 (>)
Longitud media de camino	7.4032	2.084433 (<)



(a) Grafo aleatorio de referencia (b) Grafo estelar bajo análisis

Figura 1. Comparativa de grafos para análisis

4. CONCLUSIONES

En vista a los resultados aquí expuestos, podemos afirmar que la red estelar, conformada por estrellas con similar movimiento propio, dentro del conjunto de datos de la Nube Pequeña de Magallanes, tiene una estructura que se condice con los criterios de una "red de mundo pequeño", y por consiguiente se pueden utilizar sobre la misma algoritmos de detección de comunidades y otras analíticas asociadas al estudio de redes sociales, por tratarse estas últimas también de "redes de mundo pequeño".

Estos resultados amplían en gran medida los métodos a utilizar para la detección de agrupaciones galácticas, las cuales hasta el momento se han analizado preponderantemente por medio de métodos numéricos tradicionales. La ampliación del catálogo de algoritmos a utilizar en el marco de la detección automatizada de clusters estelares es una perspectiva prometedora, considerando la velocidad de crecimiento que tienen los proyectos SURVEY a la hora de generar información digital de objetos estelares observados.

En vista de los presentes resultados, se propone avanzar en la elaboración de una técnica estandarizada de análisis de características estelares, la cual debería incluir tanto la determinación de atributos, el preprocesamiento de los mismos y la creación del grafo estelar asociado, como la caracterización final de dicho grafo como etapa previa a la selección de algoritmos de detección de comunidades (clusters).

Como parte de los trabajos futuros a desarrollarse en el marco de la tesis de doctorado, se ampliará el análisis a otros atributos, además del movimiento propio, tales como la emisión de frecuencias de los cuerpos, ya sea en espectro visible o infrarrojo, la masa, el brillo y otras características a fin de determinar si los resultados obtenidos son extrapolables a otros atributos o son particulares del movimiento propio.

Se propondrá, asimismo, algún mecanismo optimizado para el cálculo de los elementos relacionados a un objeto estelar dado, considerando que dicha operación es común a todos los análisis mencionados. El uso de técnicas de procesamiento paralelo, y optimización por medio de GPU (Graphical Processing Unit, Unidad de Procesamiento de Gráficos) tales como la tecnología CUDA de NVIDIA, ROCm de AMD o soluciones más generales como OpenCL.

REFERENCIAS

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

Barnes, J. A. y Harary, F. (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.

Cloud Computing and Big Data (s.f.). Cham, Switzerland: Springer International Publishing. isbn: 978-3-030-27713-0. url: https://link.springer.com/book/10.1007/978-3-030-27713-0.

Fall, S. M. y Chandar, R. (2012). «Similarities in populations of star clusters». En: *The Astrophysical Journal* 752.2, pág. 96.

Karttunen, H. et al. (2007). *Fundamental astronomy*. Springer.

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

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

Kumar, P., Chawla, P. y Rana, A. (2018). «A review on community detection algorithms in social networks». En: *2018 4th International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, págs. 304-309.

Kushwah, A. K. S. y Manjhvar, A. K. (2016). «A review on link prediction in social network». En: *International Journal of Grid and Distributed Computing* 9.2, págs. 43-50.

Lang, K. R. y Lang, K. R. (2013). *Essential astrophysics*. Springer.

Muñoz, R. M. et al. (s.f.). «Criterios para el diseño de una base de datos cienciométrica». En: ().

Newman, M., Barabási, A.-L. y Watts, D. J (2011). *The structure and dynamics of networks*. Princeton university press.

Paz Menvielle, M. A. et al. (2018). «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)*.

Rodríguez, M. 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.

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

Wang, C. 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, D. B. et al. (2001). *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River.