Boletim da Sociedade Astronômica Brasileira, **37**, no. 1, 110-111

Search for substructures in young stellar associations

E. Batista[1] & J. Gregorio-Hetem[1]

[1] Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo
e-mail: eduardobatista77@usp.br, gregorio-hetem@usp.br

Abstract. This work details a systematic methodology for optimizing density-based *Machine Learning* algorithms, specifically DBSCAN, for the robust identification and characterization of young stellar associations. Due to the critical sensitivity of these algorithms to input parameters (ϵ, $\mathrm{min_{PTS}}$), our approach introduces an innovative combination of analytical tools. First, a comprehensive parametric analysis is performed using two-dimensional parametric maps and quality metrics (*Silhouette Score* and *Modified Silhouette Score*). This analysis enabled the optimization focus to shift from individual parameter pairs to the intrinsic density of structures, where each optimal density is represented by an equivalence curve ($\mathrm{min_{PTS}} \propto \epsilon^5$). Second, astrometric uncertainties from *Gaia DR3* are statistically incorporated through a Bootstrap technique (with 100,000 iterations), which defines the membership probability (P) for each star. The methodology was successfully applied to the complex Canis Major OB1/R1 (CMa) association, resulting in five cohesive structures. The results demonstrate the method's validity by confirming four known structures and resolving a new substructure (Cluster 2). The final characterization included the determination of astrometric parameters, ages (all below 5 Myr), and the identification of 76 young stars with circumstellar disks.

Resumo. Este trabalho detalha uma metodologia sistemática para a otimização de algoritmos de *Machine Learning* baseados em densidade, nomeadamente DBSCAN, para a identificação e caracterização robusta de associações estelares jovens. Reconhecendo a sensibilidade crítica destes algoritmos aos parâmetros de entrada (ϵ, $\mathrm{min_{PTS}}$), nossa abordagem introduz uma combinação inovadora de ferramentas analíticas. Primeiramente, é realizada uma análise paramétrica compreensiva utilizando mapas paramétricos bidimensionais e métricas de qualidade (*Silhouette Score* e *Modified Silhouette Score*). Essa análise permitiu que o foco da otimização migrasse dos pares de parâmetros individuais para a densidade intrínseca das estruturas, onde cada densidade ótima é representada por uma curva de equivalência ($\mathrm{min_{PTS}} \propto \epsilon^5$). Em segundo lugar, as incertezas astrométricas do *Gaia DR3* são estatisticamente incorporadas por meio de uma técnica de Bootstrap (com 100.000 iterações), que define a probabilidade de pertinência (P) para cada estrela. A metodologia foi aplicada com sucesso na complexa associação Canis Major OB1/R1 (CMa), resultando em cinco estruturas coesas. Os resultados demonstram a validade do método ao confirmar quatro estruturas conhecidas e resolver uma nova subestrutura (Aglomerado 2). A caracterização final incluiu a determinação de parâmetros astrométricos, idades (todas inferiores a 5 Myr) e a identificação de 76 estrelas jovens com discos circunstelares.

Keywords. Stars: formation – Stars: pre-main sequence – Stars: evolution

1. Introduction

The identification of young stellar clusters is fundamental for the study of star and planet formation in the Galaxy. The astrometric and photometric catalog from the *Gaia* mission, especially Data Release 3 (DR3), revolutionized this field by providing data with unprecedented precision in five astrometric dimensions: position (α, δ), proper motion (μ_α, μ_δ) and parallax (ϖ). However, the search for cohesive structures in multidimensional spaces renders visual methods inefficient, requiring the use of Machine Learning algorithms. In this context, density-based methods, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise, Ester et al. 1996), have become highly effective tools for detecting stellar overdensities.

However, the effectiveness of DBSCAN is intrinsically linked to the choice of its two fundamental parameters: the neighborhood radius (ϵ) and the minimum number of points within this radius to form a core ($\mathrm{min_{PTS}}$). Even small variations in these parameters can lead to significant changes in the identification of structures, resulting in unreliable classifications that are contingent on subjective choices. Given this limitation, the literature lacks objective and systematic criteria for the optimal selection of these parameters, which hinders the reproducibility and reliability of the results. This work proposes a systematic and statistically robust methodology for optimizing these algorithms, aiming to identify the most pertinent overdensities that reflect the true physical structure of the data. We analysed the region of the Canis Major OB1/R1 (CMa) association, a complex star-forming area known for its multiple groups and subgroups, in order to validate the proposed method.

2. Optimization Methodology

The methodology is designed to circumvent the parametric sensitivity of clustering algorithms while rigorously incorporating the observational uncertainties inherent to the Gaia DR3 data. The first step involves a sweep of the parameter space ($\epsilon, \mathrm{min_{PTS}}$), using a 5D astrometric space for each combination. To guide the optimal parameter selection, we calculate three main performance metrics: the total number of clusters identified; cluster quality, evaluated by the Silhouette Score (SS), which measures internal cohesion, and by the Modified Silhouette Score (MSS), which assesses the dispersion of the cluster relative to field stars; and the number of objects belonging to the largest cluster obtained.

This parametric sweep allowed us to map the combinations that yield clusters with the best internal cohesion, separation from background objects, and best agreement with results from the literature. Our analysis revealed two distinct density profiles, associated with physical substructures in CMa: dense clusters and sparser structures. Additionally, we identified a relationship between the DBSCAN parameters whereby, for a fixed cluster density, all parameters satisfying the curve $\mathrm{min_{PTS}} \propto \epsilon^5$ result in

the identification of the same cluster, establishing an equivalence criterion.

To ensure the reliability of cluster members, astrometric uncertainty is incorporated via a Bootstrap procedure with 100,000 iterations. In each iteration, the five astrometric parameters of each star are randomly perturbed based on their uncertainties reported by Gaia. The optimized algorithm is re-executed on each perturbed sample, and the probability of a star being a cluster member (P) is calculated from its frequency of assignment to the group. Only stars with a membership probability $P \geq 50\%$ are considered reliable members, minimizing the risk of contamination by field stars.

3. Results and Discussion

The application of our methodology to the CMa region resulted in the robust identification of five cohesive stellar structures (Fig. 1), demonstrating its effectiveness. Four of these clusters (Clusters 1, 3, 4, and 5) exhibited strong agreement with structures previously cataloged in the literature, according to studies by Cantat-Gaudin & Anders (2020) and Santos-Silva et al. (2021), validating the precision of our approach. Additionally, we identified a new substructure, designated as Cluster 2, which was not previously reported in the literature. This finding confirms the validity of the developed parametric optimization and the method's ability to resolve the internal complexity of dense star-forming environments. We subsequently characterized the identified members. The kinematic and photometric ages of all clusters were estimated (Bressan et al. 2012), confirming the region's youth, with ages below 5 Myr. The analysis of color-color diagrams (Fig. 2), using data from 2MASS and AllWISE, enabled us to identify 76 objects with infrared excess emission (Class II), indicating the presence of young stars that still retain their circumstellar disks (Koenig & Leisawitz 2014).

4. Conclusions

We have established a parameter optimization for applying DBSCAN algorithms to the search for stellar clusters. The combination of a systematic parametric sweep with Bootstrap uncertainty incorporation proved to be effective for the robust detection of young stellar clusters. Its successful application to the CMa region, allowing not only the recovery of known structures but also the identification of new substructures, demonstrates the quality of the proposed method. Future work will involve expanding this methodology to other Galactic star-forming regions and developing Machine Learning models for the automated prediction of optimal parameters. This aims to maximize efficiency and accelerate the analysis of large volumes of astrometric data.

5. Acknowledgements

This study was financed by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2024/23573-0.

References

Bressan, A., Marigot, P., Girardi, L., Salasnich, B., Dal Cero, J., Rubele, S., Nanni, A., 2012, MNRAS, 427, 127.
Cantat-Gaudin, T. & Anders, F., 2020, A&A, 633, A99.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996, in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), AAAI Press, 226–231.
Koenig X. P. & Leisawitz D. T., 2014, ApJ, 791, 131
Santos-Silva, T., Gregorio-Hetem, J., Montmerle, T., Fernandes, B., 2021, MNRAS, 508, 1033.

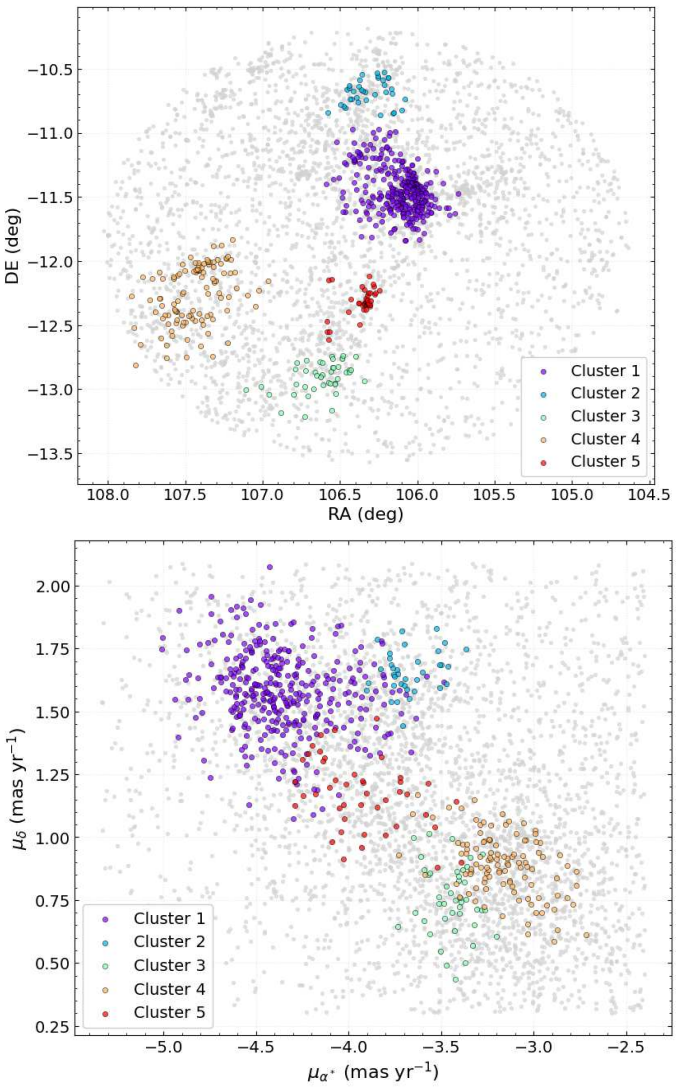

Figure 1. Identified clusters: Angular position (Top) and Proper motion (Bottom).

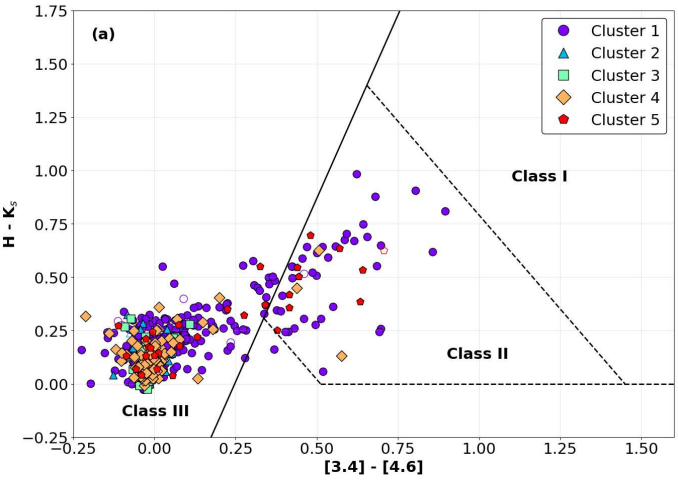

Figure 2. Color-color diagram showing the classification of YSOs combining 2MASS and WISE data. Classification regions follow the criteria of Koenig & Leisawitz (2014). Objects plotted with filled colors have good photometric quality (flags A in all bands used).