Análise de Eficiência de Algoritmos na Identificação de Aglomerados de Estrelas

Ailton Eloi Feller[1], Rodrigo Lyra[1], Rudimar Luís Scaranto Dazzi[1]

[1] Laboratório de Inteligência Aplicada
Centro de Ciências Tecnológicas da Terra e do Mar
Universidade do Vale do Itajaí (UNIVALI) – Itajaí, SC – Brazil

feller.ailton@edu.univali.br,{rlyra,rudimar}@univali.br

Abstract. *To obtain something close to a consistent sampling in star's study, first the identification of a star cluster, which is a set of stars with many common features, is made. The process of classification is often made from identification algorithms. This project's objective was to make a comparative study of different algorithms in order to evaluate identification quality and runtime of different proposals. This project presents results in a parameter analysis of each algorithm and a runtime and quality detection comparison of analyzed samples. The results showed a runtime below the other for the Star Count algorithm, but a higher quality results for K-Nearest Neighbor in generated samples.*

Resumo. *Para conseguir algo próximo de uma amostragem consistente no estudo de estrelas, é feita a identificação de aglomerados estelares, um conjunto de estrelas com muitas características em comum. O processo de identificação é muitas vezes feito a partir de algoritmos de identificação de aglomerados. O objetivo deste trabalho foi fazer um estudo comparativo de diferentes algoritmos, com o intuito de avaliar diferentes propostas quanto a qualidade de identificação e o seu tempo de execução. Este trabalho traz como resultado uma análise dos parâmetros de cada algoritmo e um comparativo do tempo de execução e qualidade de detecção das amostras analisadas. Os resultados mostraram um tempo de execução abaixo dos demais para o algoritmo Star Count, mas uma qualidade de resultados maior para o K-Nearest Neighbor nas amostras geradas.*

1. Introdução

Estrelas são corpos celestes que emitem luz, são compostas principalmente por hidrogênio e hélio, mas também por materiais mais pesados como ferro e carbono. Elas nascem a partir de nuvens de gás de hidrogênio molecular, densas e muito frias. Elas passam a maior parte de sua vida fazendo fusão nuclear, o que gera luz, calor e transforma hidrogênio em hélio. Quando chega ao fim de sua vida, com o esgotamento do hidrogênio, dependendo de seu tamanho, ela pode fundir elementos mais pesados até explodir em uma supernova, ou se transformar em uma anã branca ou estrela de nêutrons. Esse processo despeja material com elementos mais pesados no espaço que se acumulam para gerar uma nova geração de estrelas. A origem dos planetas e mesmo da vida são

provenientes deste processo, e o estudo da evolução das estrelas é importante para entendermos melhor o seu funcionamento e suas implicações (OBSERVATÓRIO NACIONAL, 2013).

As estrelas preferencialmente são formadas em aglomerados, não em isolamento. Os aglomerados surgem a partir de nuvens moleculares gigantes e permanecem inseridos nelas nos estágios iniciais de sua formação. As estrelas formadas nesse processo têm, geralmente, a mesma idade e possuem distância relativa entre si desprezíveis em relação a um observador na Terra. No estágio final, o aglomerado emerge da nuvem podendo então permanecer como uma entidade unida ou se desmembrar (OLIVEIRA, 2004).

O objetivo da pesquisa é analisar a eficiência de algoritmos, previamente selecionados de acordo com o trabalho de Schmeja (2011), para a identificação de aglomerados estelares com base na densidade. Os algoritmos a serem implementados serão aplicados em apenas uma linguagem de programação buscando sempre a forma mais abstraída possível do problema, visando a melhor performance e confiabilidade dos resultados. Depois de implementados os algoritmos serão testados com ambientes de testes controlados, criados a partir da geração de instâncias aleatórias seguindo padrões baseados no trabalho de Schmeja (2011).

Este artigo traz o estudo feito e os resultados obtidos a partir da aplicação dos algoritmos escolhidos para esse trabalho: Star Count, KNN, Voronoi Tesselation, MST e DBScan, junto com análises e conclusões sobre os desempenhos destes algoritmos.

2. Desenvolvimento

Diferente de outras áreas de pesquisa, o estudo de evolução estelar não tem realizado esforços para uma amostragem controlada em laboratório para suas análises. Os aglomerados de estrelas, por consistirem de estrelas que nasceram em um espaço de tempo relativamente curto e que vieram da mesma nuvem de gás molecular são uma forma de conseguir uma amostragem mais homogênea (THAN, 2006).

Os trabalhos desta área utilizam, em sua grande maioria, algoritmos para a identificação inicial dos aglomerados em seus estudos. Muitos deles não fazem uma análise para a escolha do algoritmo utilizado, ou não mostram atenção a isso em suas publicações. Os algoritmos mais utilizados são o MST e o KNN, com exceções que utilizam de outras ferramentas como o Voronoi Tesselation (ESPINOZA, SELMAN e MELNICK 2009) ou o Star Count (SCHMEJA 2011).

O estudo mais próximo do objetivo da pesquisa é o de Schmeja (2011), onde ele traz quatro algoritmos já usados na literatura com o fim de reconhecer aglomerados estelares e faz um comparativo entre eles. Ele executa esses algoritmos em ambientes de testes controlados gerados aleatoriamente utilizando padrões conhecidos de aglomerados de estrelas abertos e globulares. O maior foco do estudo de Schmeja é com o reconhecimento, tanto parcial quanto total, das estrelas pertencentes ao aglomerado, não avaliando o seu tempo de execução.

2.1. Gerador de Amostras

Para esse trabalho foi criado um gerador de amostras em um ambiente de testes controlado, usando o conhecimento dos tipos de aglomerados estelares globulares e

usando uma modificação da fórmula da transformada de Box-Müller (BOX; MÜLLER, 1958).

As equações 1 e 2 foram feitas para esse trabalho a partir da fórmula da transformada de Box-Müller, elas geram a posição de um ponto em um plano cartesiano começando das coordenadas 0,0 e com o afastamento partindo de uma distribuição normal. A primeira parte das equações diferem por **A** e **B**, que representam as proporções dos eixos, valores diferentes formam um aglomerado no formato de uma elipse, enquanto valor iguais formam um aglomerado no formato de um círculo. Dentro da raiz quadrada temos a geração do afastamento em relação ao ponto inicial com **u1** sendo um número real aleatório entre 0 e 1 e **variância** representando a dispersão dos objetos. O próximo passo da equação determina o ângulo de deslocamento utilizando cosseno para determinar o deslocamento no eixo X e seno para o deslocamento no eixo Y, a variável **u2** é também um número real aleatório entre 0 e 1 e é multiplicado por 2π para gerar o ângulo em radianos.

A última parte da equação é destinada aos aglomerados em formato de anel, se o **raio** é 0, essa parte não soma nada a equação e o aglomerado fica globular. Caso contrário, o ponto de origem da posição 0,0 é deslocado para eixo X e eixo Y de forma similar a distribuição, com o **raio** sendo a distância do centro, e **u3** uma variável aleatória que determina o ângulo.

$$pos\ X = A * \sqrt{\ln(u1) * -2 * variância}\ * \cos(u2 * 2\pi) + raio * \cos(u3 * 2\pi) \qquad (1)$$

$$pos\ Y = B * \sqrt{\ln(u1) * -2 * variância}\ * \operatorname{sen}(u2 * 2\pi) + raio * \operatorname{sen}(u3 * 2\pi) \qquad (2)$$

A partir das diferentes formas de aglomerados e de sua composição, foram geradas diferentes amostras de modo a abranger o cruzamento de todas as seguintes categorias:

- Formato: globular (raio = 0) e anel (raio = 1);
- Densidade: alta (1), média (0.5) e baixa (0.1) variância;
- Proporção dos eixos: globo (A = B), elipse (A = 2B);
- Quantidade de estrelas: alta (500), média (200) e baixa (50);
- Quantidade de estrelas no ruído de fundo (Em relação a quantidade de estrelas do aglomerado): alta (5x), média (3x), baixa (1x).

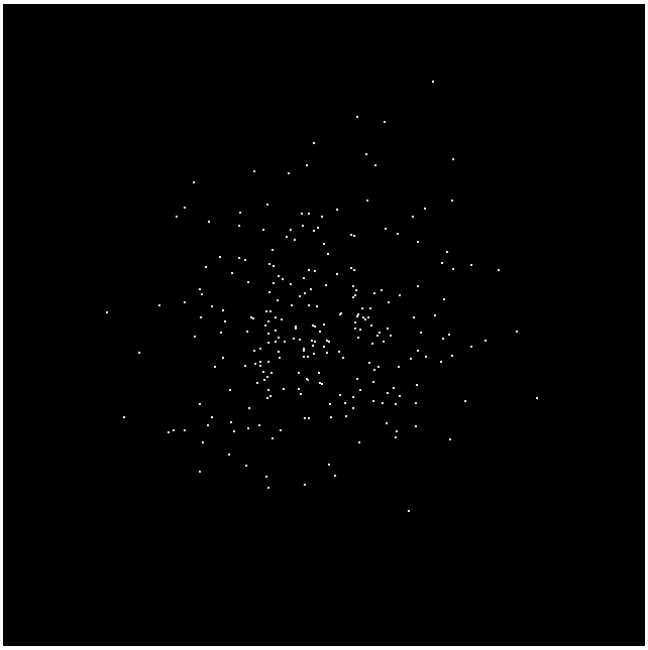

Figura 1. Fórmula aplicada em um aglomerado globular

A figura 1 traz uma referência visual da fórmula aplicada utilizando o raio como 2, para gerar um aglomerado em forma de anel, A como 2 e B como 1 para formar um disco e a variância em 0,1 para uma distribuição mais compacta.

Além das fórmulas descritas acima, para a geração dos padrões de aglomerados, há um gerador de pontos aleatórios sem parâmetros para a criação de estrelas que não pertencem ao aglomerado, simulando uma situação real onde nem todas as estrelas presentes na imagem pertencem ao aglomerado.

2.2. Algoritmos

2.2.1. Star Count

O algoritmo Star Count é recorrente nos trabalhos de detecção de aglomerados estelares (SCHMEJA, 2011), (SCHMEJA, 2014), (KIM; JERJEN, 2015), funciona de forma lúdica ao nome, dividindo um plano bidimensional em vários quadrantes e fazendo a contagem das estrelas em cada quadrante. A partir daí cada quadrante tem seu número de estrelas comparado a um limite que define se o quadrante faz parte de um aglomerado. Quadrantes adjacentes que excedem ao limite são considerados como parte do mesmo aglomerado. Star Count é o nome do algoritmo em alguns trabalhos, incluindo o trabalho principal pesquisado, em outros aparece com diferentes nomenclaturas, mas todos usam o mesmo princípio de aplicação.

2.2.2. KNN (K-Nearest Neighbor)

O algoritmo KNN (K-Nearest Neighbor ou K-Vizinho Mais Próximo) é recorrente na literatura para a identificação de aglomerado (SCHMEJA; KUMAR; FERREIRA, 2008), (SCHMEJA, 2011), (GOULIERMIS et al., 2003). É usado como um avaliador de

densidade pela distância que um objeto tem em relação ao seu K vizinho mais próximo. Uma analise verifica a distância de um objeto com todos seus vizinhos, as ordena e retira a informação que fica na posição K, a probabilidade deste objeto pertencer a um aglomerado é inversamente proporcional à distância encontrada.

2.2.3. Voronoi Tessalation

O Voronoi Tessalation não é necessariamente um algoritmo, mas uma representação matemática gerada por um, e é usado em trabalhos relacionados (CARTWRIGHT; MOSS; CARTWRIGHT, 2011), (SCHMEJA, 2011). Essa abordagem parte da divisão de um espaço euclidiano em células ou regiões. Para esta tarefa cada região é gerada a partir de um ponto gerador, tal que cada segmento pertencente que delimita esta área não pode estar mais próximo de nenhum outro ponto do plano que não seja o gerador desta região.

O algoritmo utilizado para criação do diagrama de Voronoi foi escolhido de acordo com a sua complexidade, com base nesse critério, foi escolhido o algoritmo de Fortune (FORTUNE, 1987). A figura 2 mostra a aplicação do algoritmo, a parte colorida são os aglomerados, enquanto as linhas brancas apenas mostram os segmentos dos polígonos gerados pelos pontos restantes e que não fazem parte de nenhum aglomerado.

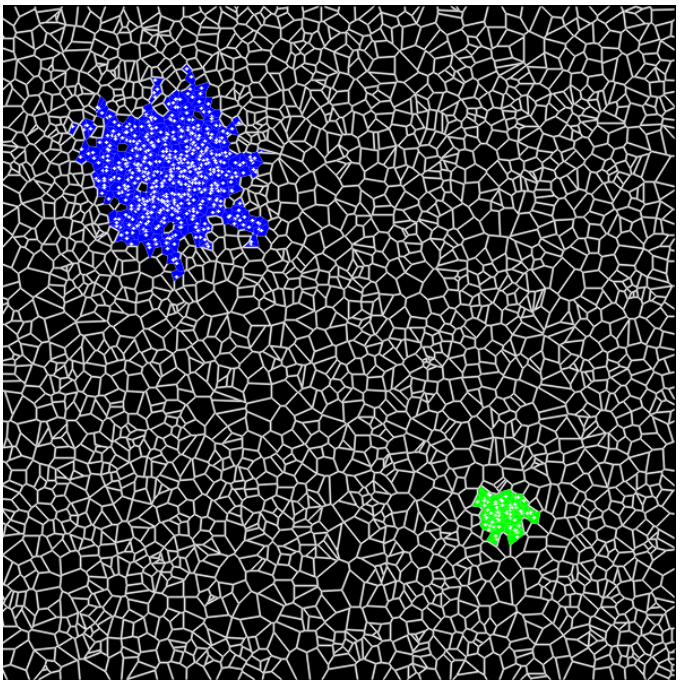

Figura 2. Amostra com dois aglomerados após a aplicação do algoritmo.

2.2.4. MST (Minimum Spanning Tree)

O algoritmo MST (Minimum Spanning Tree ou Arvore Geradora Minima) é o mais recorrente na literatura para a identificação de aglomerado (BASTIAN et al., 2007), (SCHMEJA, 2011), (SCHMEJA; KUMAR; FERREIRA, 2008). Esse algoritmo forma um grafo usando a distância entre as estrelas no espaço de busca, então é gerada uma árvore que une todos os objetos usando a menor soma das distâncias possíveis. Um passo

adicional para criar essa arvore é uma poda posterior a geração da árvore inicial. Essa poda elimina os maiores ramos, mantendo apenas subárvores que possuem ligações mais próximas entre si.

2.2.5. DBScan

O algoritmo DBScan, mesmo não sendo utilizado nos estudos com aglomerados estelares, ele é recorrente em trabalhos computacionais (DUDIK et al., 2015). O algoritmo avalia cada ponto no espaço e todos os vizinhos próximos dentro de uma distância D dele. Se a contagem dos vizinhos nessa área for maior que o limiar determinado, este é classificado como um *core point* e pertencente ao aglomerado. Um aglomerado é formado por todos os pontos classificados como *core point* que são vizinhos e todos os pontos que não foram classificados, mas estão dentro do raio de distância D.

3. Resultados

A Figura 3 representa a variação do tempo em relação ao tamanho total da amostra em um comparativo entre os algoritmos. Neste gráfico o nível de ruído e número de estrelas dos aglomerados foram calculados em uma única variável do número total de estrelas para representar melhor a complexidade de tempo das instâncias. A escala de tempo em segundos é logarítmica para representar melhor a diferença de tempo entre alguns algoritmos nas amostras maiores.

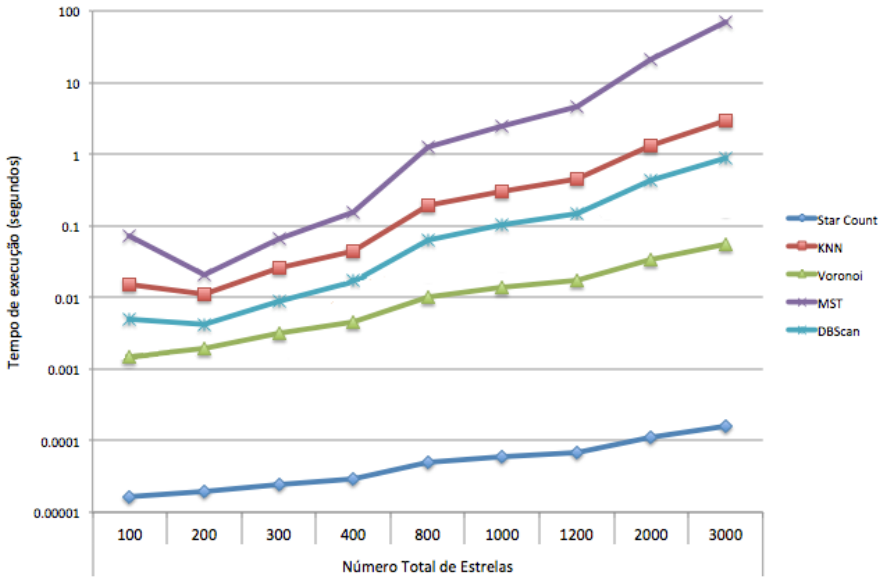

Figura 3. Gráfico comparativo de tempo entre os algoritmos

A Figura 4 mostra um comparativo da média de qualidade em relação aos tamanhos de amostras entre os algoritmos. O algoritmo KNN traz as melhores médias de resultado e se mantém estável com uma pequena queda nas amostras menores. Star Count e DBScan ficam com a qualidade um pouco abaixo da média e muito parecidas entre si, com uma desvantagem para o Star Count em amostras com o nível de ruído muito alto.

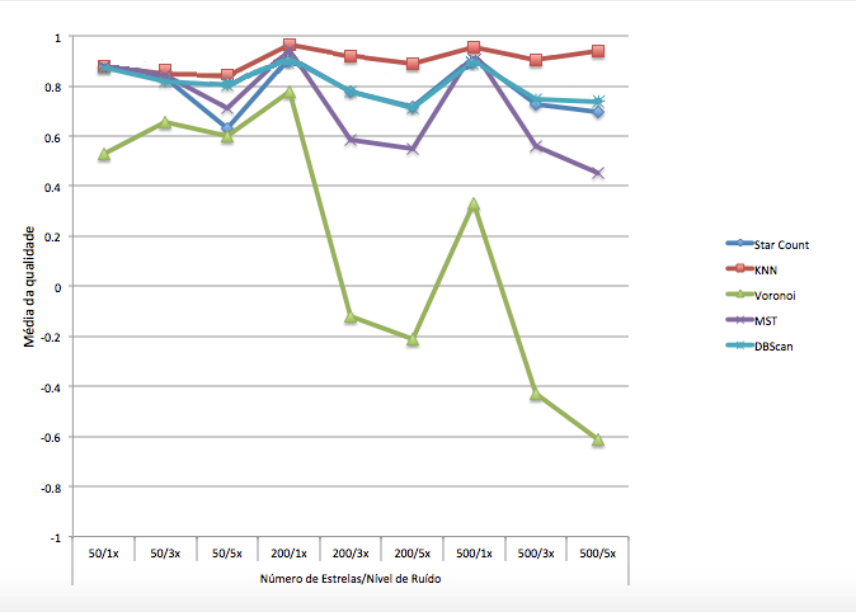

Figura 4. Gráfico comparativo da qualidade dos resultados de cada algoritmo

A figura 5 mostra um gráfico fazendo a comparação do impacto das proporções dos eixos dos aglomerados na qualidade do resultado de cada algoritmo, a figura 6 mostra a comparação do impacto da densidade de estrelas na qualidade de resultado em cada algoritmo e a figura 7 mostra a comparação do impacto do tamanho do raio na qualidade de resultado em cada algoritmo. Os gráficos não possuem um valor fixo na escala vertical, apenas mostra a relação entre os resultados de cada algoritmo.

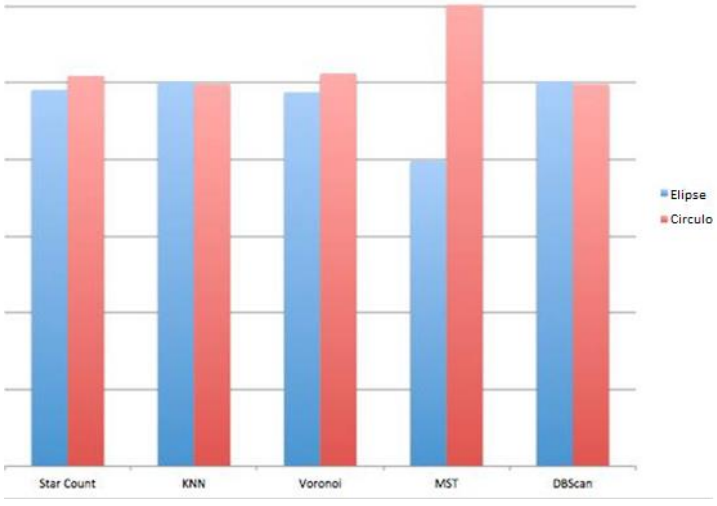

Figura 5. Comparação entre a proporção dos eixos.

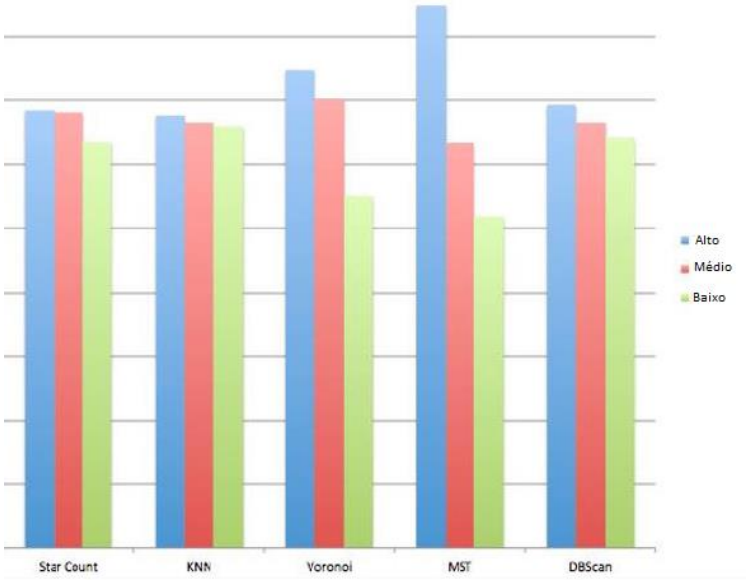

Figura 6. Comparação entre as densidades.

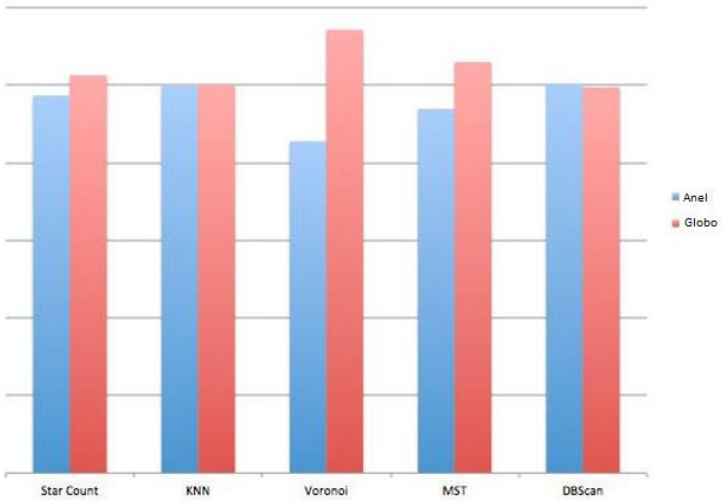

Figura 7. Comparação entre os raios.

4. Conclusões

Depois dos testes com as 5 alternativas serem finalizados, o KNN se provou o melhor algoritmo para resolver o problema, apesar do seu tempo de execução ter sido o segundo pior dentre os algoritmos, ele possui a melhor média de qualidade nos aglomerados. Para as amostras testadas o valor do parâmetro K foi definido para 16, e baseado nos trabalhos que também utilizam esse tipo de algoritmo, a tendência para melhores resultados está próximo dessa faixa, onde não sofre muito das flutuações de ruído e também não perde a capacidade de detectar os elementos da borda.

O algoritmo mais rápido dentre todos foi o Star Count, que trouxe uma diferença visualmente grande pela escala logarítmica do gráfico de tempo, mas que a tendência seria se afastar mais em relação aos outros à medida que o tamanho da amostra aumenta. A união deste algoritmo com o KNN foi observada como uma possibilidade promissora,

aliando o baixo tempo de execução do Star Count para diminuir o tamanho da amostra e a melhora na qualidade dos testes ao utilizar o KNN tentando diminuir seu tempo de execução.

O algoritmo MST mostrou um crescimento de tempo muito grande, ficando muito além dos demais, o seu crescimento de tempo atrapalharia testes consecutivos em espaços de busca com muitas estrelas. Apesar de mostrar resultados de qualidade regulares para as amostras geradas, o KNN ainda obteve melhores resultados. Outro problema deste algoritmo é sua tendência a falsos positivos em grupos muito grandes sem aglomerados, com uma média mais homogênea ele acaba deixando muitas sub árvores que acabam minimizando seus resultados.

O algoritmo baseado no diagrama de Voronoi apresentou diversos problema no decorrer da execução deste projeto, os resultados iniciais com ele foram mais inconsistentes do que os apresentados neste documento, a medida de qualidade foi modificada de tamanho de segmento por tamanho de área e o algoritmo foi reescrito, mas apesar dos resultados melhorarem ele ainda apresenta muitos problemas e anomalias. Ele aparenta ser muito sensível ao ruído, que pode causar uma diminuição no tamanho de áreas fora do aglomerado, a disposição dos vizinhos também interfere neste parâmetro que pode trazer uma área muito pequena ou muito grande, que interfere no tamanho geral. As bordas laterais também geravam um aumento geral da média de áreas por só terem pares geradores em alguns sentidos, problema este já solucionado na etapa de desenvolvimento. Os resultados se apresentam muito abaixo dos demais, ou com picos que não apresentam muita confiabilidade. Apesar de todos os esforços, não é possível descartar uma falha de implementação no algoritmo.

Referências

BOX, G. E. P.; MULLER, Mervin E.. A Note on the Generation of Random Normal Deviates. **Ann. Math. Statist.,** [s.l.], v. 29, n. 2, p.610-611, jun. 1958. Institute of Mathematical Statistics. DOI: 10.1214/aoms/1177706645.

BASTIAN, N. et al. Hierarchical star formation in M33: fundamental properties of the star-forming regions. **Monthly Notices Of The Royal Astronomical Society,** [s.l.], v. 379, n. 4, p.1302-1312, 21 ago. 2007. Oxford University Press (OUP). DOI: 10.1111/j.1365-2966.2007.12064.x. Disponível em: <http://arxiv.org/pdf/0706.0495v1.pdf>. Acesso em: 10 dez. 2016.

CARTWRIGHT, Annabel; MOSS, Jennifer; CARTWRIGHT, Joe. New statistical methods for investigating submarine pockmarks. **Computers & Geosciences,** [s.l.], v. 37, n. 10, p.1595-1601, out. 2011. Elsevier BV. DOI: 10.1016/j.cageo.2011.02.013. Disponível em: <http://api.elsevier.com/content/article/PII:S0098300411000884?httpAccept=text/xml>. Acesso em: 10 dez. 2016.

DUDIK, Joshua M. et al. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. **Computers In Biology And Medicine,** [s.l.], v. 59, n. 1, p.10-

18, abr. 2015. Elsevier BV. DOI: 10.1016/j.compbiomed.2015.01.007. Disponível em: <http://api.elsevier.com/content/article/PII:S0010482515000244?httpAccept=text/xml>. Acesso em: 10 dez. 2016.

ESPINOZA, P.; SELMAN, F. J.; MELNICK, J.. The massive star initial mass function of the Arches cluster. Astronomy & Astrophysics, [s. I.], v. 501, n. 4, p.563-583, abr. 2009. Mensal. Disponível em: <http://cds.aanda.org/articles/aa/pdf/2009/26/aa8597-07.pdf>. Acesso em: 10 dez. 2016.

FORTUNE, S. A sweepline algorithm for Voronoi diagrams. Algorithmica, v. 2, n. 1-4, p. 153–174, nov. 1987.

GOULIERMIS, Dimitrios A. et al. HIERARCHICAL STELLAR STRUCTURES IN THE LOCAL GROUP DWARF GALAXY NGC 6822. **Apj,** [s.l.], v. 725, n. 2, p.1717-1734, 1 dez. 2010. IOP Publishing. DOI: 10.1088/0004-637x/725/2/1717. Disponível em: <http://arxiv.org/pdf/1010.1940v1.pdf>. Acesso em: 10 dez. 2016.

KIM, Dongwon; JERJEN, Helmut. A HERO'S LITTLE HORSE: DISCOVERY OF A DISSOLVING STAR CLUSTER IN PEGASU.**The Astrophysical Journal,** U.s.a, v. 799, n. 1, p.73-81, 20 jan. 2015. Disponível em: <http://iopscience.iop.org/0004-637X/799/1/73/pdf/0004-637X_799_1_73.pdf>. Acesso em: 10 dez. 2016.

OBSERVATÓRIO NACIONAL. EAD 2013: Astrofísica Geral. 2013. Disponível em: <http://www.on.br/ead_2013/>. Acesso em: 10 dez. 2016.

OLIVEIRA FILHO, K. de S.; SARAIVA, M. de F. O.. Astronomia e Astrofísica. 2. ed. São Paulo: Livraria da Física, 2004.

SCHMEJA, S.. Identifying star clusters in a field: A comparison of different algorithms. Astronomische Nachrichten, Heidelberg, Alemanha, v. 332, n. 2, p.172-182, fev. 2011. Mensal. Disponível em: <http://arxiv.org/pdf/1011.5533v1.pdf>. Acesso em: 10 dez. 2016.

SCHMEJA, S. et al. Global survey of star clusters in the Milky Way: III. 139 new open clusters at high Galactic latitudes.**Astronomy & Astrophysics,** Eso, v. 566, n. 1, p.1-11, 24 jun. 2014. Mensal. Disponível em: <http://arxiv.org/pdf/1406.6267v1.pdf`>. Acesso em: 10 dez. 2016.

SCHMEJAc, S.; KUMAR, M. S. N.; FERREIRA, B.. The structures of embedded clusters in the Perseus, Serpens and Ophiuchus molecular clouds. **Monthly Notices Of The Royal Astronomical Society,** Rsa, v. 389, n. 1, p.1209-1219, jan. 2008. Disponível em: <http://arxiv.org/pdf/0805.2049v1.pdf>. Acesso em: 10 dez. 2016.

THAN, K.. Star Clusters Hold Secrets to Stellar Evolution. 2006. Disponível em: <http://www.space.com/2306-star-clusters-hold-secrets-stellar-evolution.html>. Acesso em: 10 dez. 2016.