Programa Computacional para a Identificação de Objetos Celestes por meio da Análise de Curvas de Luz

PALMA, Wallace Pannace[1]
ZALEWSKI, Willian[2]

RESUMO

Estrelas variáveis são uma das principais fontes de informação para a astronomia, o estudo da variação do brilho dado em curvas de luz, entre outros dados, pode fornecer informações acerca do seu sistema, como se possui exoplanetas. Neste trabalho, analisamos 36 sistemas com exoplanetas identificados de observações feitas pelo telescópio espacial NASA Kepler. Nesta análise buscamos por padrões na variação do brilho que se repitam nestes sistemas, que possam estar relacionados com a existência dos exoplanetas. Os padrões buscados são padrões morfológicos presentes nas curvas de luz de cada sistema, os dados são analisados em formas de séries temporais, utilizando técnicas para encontrar motifs, possibilitando análises temporais do fenômeno.
Palavras-chaves: astronomia, séries temporais, clusterização, aprendizado de máquina.

1 INTRODUÇÃO

Estrelas são gigantes esferas de gás que estão realizando um delicado equilíbrio entre a força da gravidade, que tenta esmagar toda a massa em uma esfera menor no centro da estrela, e a força de queima nuclear no núcleo da estrela tentando expandir. Grande parte das estrelas que observamos possuem brilho praticamente constante com variações perceptíveis em ordem de 10 a 100 mil anos. Há também estrelas variáveis, que seu brilho apresenta variação perceptível de frações de segundos até alguns anos. A variação do brilho de uma estrela é uma das fontes mais importantes de dados astrofísicos, que possibilita obter informações sobre seu interior, massa, raio, rotação e atividade estelar, seus sistemas e descoberta de exoplanetas. Essas informações acerca do brilho das estrelas são obtidas principalmente, atualmente, por satélites, como o pioneiro CoRoT (dezembro de 2006) e como o Kepler (março de 2009), que foi um marco na pesquisa de exoplanetas, os dados de sua fotometria em geral são dados em formas de curvas de luz. Uma curva de luz é uma curva onde são representadas as variações de brilho da estrela no decorrer do tempo.

[1]Discente do curso de Engenharia Física do – ILACVN – UNILA; bolsista ITI-UNILA. E-mail: wpd.palma.2016@aluno.unila.edu.br.
[2]Docente do – ILATIT – UNILA. Orientador de bolsista ITI-UNILA. E-mail: willian.zalewski@unila.edu.br.

Com o grande avanço em tecnologia e investimentos em telescópios, cresceu a quantidade de dados, se tornando uma tarefa inviável processá-los manualmente. Neste cenário, técnicas de aprendizado de máquina e estatísticas têm se tornado importantes para o entendimento e processamento destes dados. Diversos pesquisadores têm utilizado estas técnicas para contribuir com a classificação destes dados. Nos trabalhos efetuados até então, para a classificação das curvas de luz, utiliza-se transformações como a de transformação de Fourier, Wavelet, entre outras, que passam a informação do domínio do tempo para outro, como a frequência, por consequência a informação temporal é deixada de lado. Para utilizar essa informação temporal, podem ser aplicadas técnicas de séries temporais, como ferramentas para discriminar curvas de luz através de padrões morfológicos que se repetem ao decorrer da curva (motifs).

O objetivo deste trabalho consiste em desenvolver um programa computacional para a identificação de corpos celestes por meio da busca por padrões em curvas de luz utilizando análise de séries temporais e clusterização.

2 METODOLOGIA

Para este trabalho buscamos aplicar técnicas de séries temporais em dados astronômicos, especificamente para a identificação de exoplanetas. Esta escolha se deve ao aspecto de periodicidade da translação de planetas e seus efeitos morfológicos em curvas de luz. Tendo isto em vista, a base de dados disponibilizada pela *kaggle exoplanet hunting in deep space,* apresenta séries temporais classificadas e são derivadas de observações feitas pelo telescópio espacial NASA Kepler. Outra característica interessante desta base de dados é que as observações feitas são periódicas, isto possibilita uma comparação morfológica adequada.

A identificação dos motifs é determinada, geralmente, por uma função de distância (D) e a fixação de um limiar de aceitação (r). Assim, se r é um limiar de aceitação real positivo, uma série temporal com uma subsequência $C_1\,C_1$ iniciada na posição p, e outra subsequência $C_2\ C_2$ na posição q, seja D a distância entre dois objetos, $D(C_1, C_2) \leq r\ D(C_1, C_2) \leq r$, então assume-se que a subsequência $C_1\,C_1$ é similar a subsequência $C_2\,C_2$. Neste trabalho, a distância euclidiana foi utilizada para o agrupamento das subsequências encontradas.

Como os dados de diferentes séries foram comparados, todas as curvas de luz e subsequências separadas são normalizadas para minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Para sabermos quais subsequências se repetem em uma curva de luz, foi utilizada a técnica *Matrix Profile* (MP). Nessa técnica, seja T uma curva de luz, a matrix profile se trata de uma comparação da curva de luz normalizada com ela mesma por distância euclidiana, onde a i-ésima posição marca a distância da subsequência em T, na i-ésima posição, ao seu vizinho mais próximo presente em T onde quer que este esteja na curva, possibilitando um gráfico de posição por menor distância euclidiana em toda a séria excluindo combinações triviais.

No agrupamento das subsequências selecionadas, para que cada grupo seja a representação de um motif, buscamos minimizar as distâncias intra-grupos e maximizar a distância extra-grupos. Para sabermos a qualidade do agrupamento, utilizamos a medida de qualidade para clusters chamada *Silhouette Coefficient*, que é calculado da seguinte maneira, a média das distâncias intra-grupo (a) e a distância entre a amostra e o grupo mais próximo que ela não pertence (b) para cada amostra, o *Silhouette Coefficient* é dado por $(b - a)/max(a, b)$ $(b - a)/max(a, b)$ sendo 1 o melhor valor possível para a qualidade é -1 o pior resultado.

Para o desenvolvimento desse projeto, optamos por utilizar a linguagem Python 3.0 por identificarmos ser suficiente para nossa proposta, ser gratuita, prática e ser amplamente utilizada em produções científicas. Considerando a grande quantidade de dados, o processamento dos algoritmos foi realizado com o cluster C3HPC (UFPR) que contém 6 nodos de processamento, cada um com 4 sockets Intel Xeon E5-4627 v2 @ 3.30GHz (8 núcleos por socket) e 256 GB de RAM.

3 FUNDAMENTAÇÃO TEÓRICA

O trânsito planetário ou eclipses estelares é uma maneira de inferir a existência de planetas no sistema de uma estrela, é um fenômeno acromático onde ocorrem quedas periódicas na intensidade do brilho observado da estrela, trânsitos planetários só podem ser identificados através deste método em estrelas que possuem um ou mais planetas gigantes gasosos que passe exatamente entre a estrela e o observador (no caso o satélite Kepler). Na Figura 1, podemos observar

um exemplo de trânsito planetário e seu efeito sobre a curva de luz obtida pela observação da estrela.

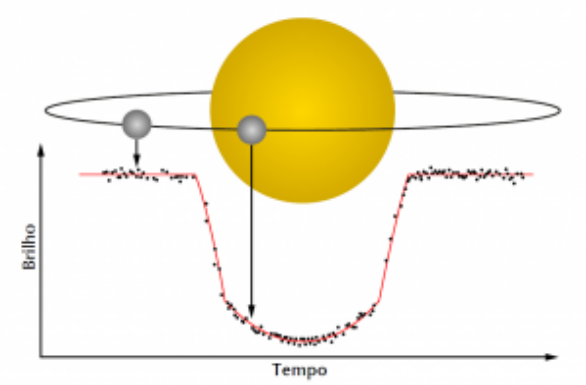

Figura 1: Curva de luz gerada pelo trânsito planetário

Uma série temporal pode ser definida como um conjunto de observações de um determinado fenômeno realizadas ao longo do tempo. Portanto uma característica importante ao se considerar a informação temporal refere-se à ordem das observações que a compõem, ou seja, busca-se analisar e modelar a dependência que uma observação vizinha possui com outra. Consideramos as curvas de luz como sendo séries temporais estacionárias (se desenvolve no tempo ao redor de uma média constante, refletindo alguma forma de equilíbrio estável).

A abordagem utilizada neste trabalho para o processamento das curvas de luz consiste na identificação de motifs para analisar as séries temporais. Um motif é basicamente um padrão frequente desconhecido em uma série temporal, o qual possui a capacidade de descrever essa série. Buscamos obter estas informações locais provindas dos motifs que podem fornecer informações relevantes para extração de conhecimento em dados temporais. Para a aquisição de motifs, utilizamos técnicas de agrupamento não supervisionado. O agrupamento é utilizado para separar integrantes baseados em suas semelhanças, no caso morfológicas, e distinções características dos dados sem a predefinição de categorias.

4 RESULTADOS

Como pode ser observado na Figura 2 (a), o melhor resultado para o valor ‘r’ para este conjunto de dados foi 0,1 no intervalo considerado, implica dizer que o agrupamento feito com este valor é provavelmente o melhor para se buscar por informações.

Neste agrupamento foram obtidos 746 grupos com o número de integrantes descritos conforme a Figura 2 (b).

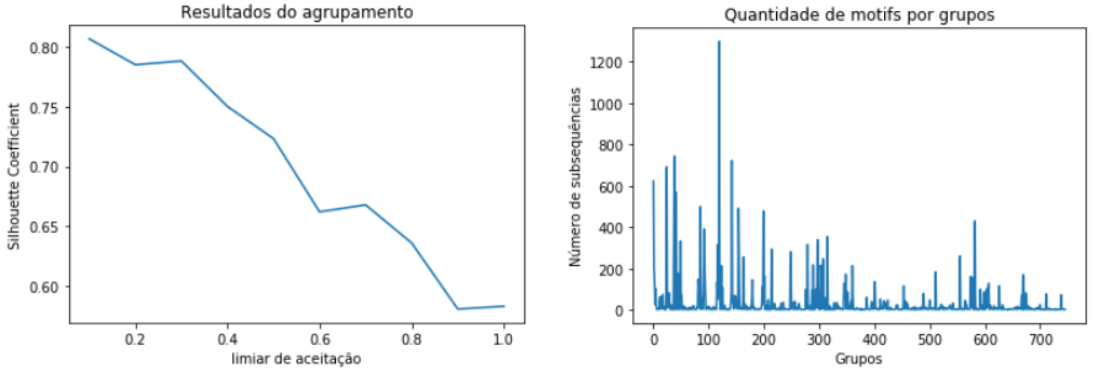

Figura 2: (a) Resultados do agrupamento; (b) Quantidade de motifs por grupos

5 CONCLUSÕES

Com a utilização de técnicas de séries temporais, implementamos um programa computacional com o qual extraímos padrões das curvas de luz. Obtivemos bons resultados na separação destes padrões, considerando a qualidade medida do agrupamento realizado. Com base nesse agrupamento, estratégias que utilizam aprendizado de máquina em combinação com algoritmos que façam contagem de quantas vezes uma subsequência ocorre em um período de tempo poderão ser utilizados para a classificação das curvas de luz, entre outras. Estes conceitos se aplicam também a outras situações com comparações morfológicas e/ou séries temporais, logo, a implementação computacional efetuada pode ser utilizada para buscar por padrões em outras bases de dados.

6 PRINCIPAIS REFERÊNCIAS BIBLIOGRÁFICAS

MITCHELL, T. M. Machine Learning. Boston, USA: McGraw-Hill, 1997.

ZALEWSKI, Willian. Modelagem Simbólica de Padrões Morfológicos para a Classificação de Séries Temporais. Curitiba, PR, p. 55-58, 2015.

The UCR Matrix Profile Page. Disponível em: <http://www.cs.ucr.edu/~eamonn/MatrixProfile.html>. Acesso em: 2018.

EHLERS, R. S. Análise de Séries Temporais. Curitiba - PR, 2005

CASTRILLÓN, J. P. B. Análise de Curvas de Luz do Corot usando diferentes processos comparativos: estimando períodos de rotação estela. UFRN, 2010.