PROGRAMA COMPUTACIONAL PARA A IDENTIFICAÇÃO AUTOMÁTICA DE EXOPLANETAS

MONTANGER, Patricia Oliveira[1]
ZALEWSKI, Willian[2]

RESUMO
A coleta de informações ao longo do tempo se aplica em inúmeras situações do mundo atual e ocorre exponencialmente, sendo de grande interesse a análise desses dados da maneira mais rápida e eficaz possível. Quando dados são coletados ao longo do tempo estes podem ser representados por meio de uma série temporal, é o que acontece com as curvas de luz de exoplanetas, que são nosso objeto de estudo. Neste trabalho buscamos o desenvolvimento de métodos para a análise inteligente de dados de séries temporais a partir da aplicação de gráficos de recorrência, que são uma ferramenta de visualização de séries temporais multivariadas baseada na exploração do comportamento recorrente característico de uma série temporal. Com isso, realizamos a identificação de exoplanetas por meio de algoritmos de aprendizagem de máquina e analisamos os resultados junto a validação cruzada que avalia o desempenho do estimador, nos permitindo identificar os melhores modelos de classificação.
Palavras-chaves: séries temporais, curvas de luz, gráficos de recorrência, aprendizado de máquina.

1 INTRODUÇÃO

Em diversas áreas do conhecimento estão presentes informações que estão sujeitas a variações temporais, como na economia com os preços diários de ações, na medicina em eletrocardiogramas, na meteorologia e na astrofísica com a identificação de objetos celestes. Estes exemplos tão importantes e comuns no dia a dia mostram como o desenvolvimento tecnológico referente ao armazenamento e ao processamento de dados temporais é importante. O tipo de dado temporal mais comum é chamado de série temporal, a qual pode ser entendida como um conjunto ordenado de observações registradas cronologicamente. Neste trabalho, a série temporal de interesse é a variação da intensidade luminosa de corpos celestes coletada pelo telescópio Kepler da NASA, o qual nos fornece uma base de dados com cadência máxima de 30 minutos entre cada registro, denominadas curvas de luz. Esse tipo de dado possibilita obter diversas informações sobre os valores de

[1] Estudante do Curso de Engenharia Física – ILACVN – UNILA; bolsista ITI-UNILA. E-mail: patricia.montanger@aluno.unila.edu.br.
[2] Doutor – ILATIT– UNILA. Orientador de bolsista ITI-UNILA; E-mail: willian.zalewski@unila.edu.br.

massa e raio de estrelas, supernovas, sistemas binários e, em especial, pode ser utilizado para identificação de exoplanetas por meio da análise do trânsito planetário.

As abordagens tradicionais para a análise de séries temporais são baseadas em métodos estatísticos, os quais, em geral, não se mostram eficientes em domínios de dados não lineares. Estes métodos analisam cada dado da série independentemente, sem considerar o fato de que existe uma relação temporal entre as observações realizadas. Mediante esta restrição das abordagens estatísticas, muitos estudos propuseram a utilização de técnicas de aprendizado de máquina [3]. Essas técnicas são baseadas na inferência indutiva, a qual possibilita derivar novos conhecimentos automaticamente a partir de outros previamente conhecidos [1]. Nesse contexto, neste trabalho estudamos as curvas de luz por meio da técnica de gráficos de recorrência em combinação com algoritmos de aprendizado de máquina. Nosso objetivo é contribuir de modo significativo para o processo de classificação automática de exoplanetas, agregando informações para auxiliar no processo de tomada de decisões de astrônomos.

2 FUNDAMENTAÇÃO TEÓRICA

Como mencionado, nosso objeto de estudo são as curvas de luz que caracterizam o trânsito planetário de exoplanetas, as quais podem ser entendidas como séries temporais constituídas pela variação do brilho de um objeto celeste no tempo. Para compreender este trabalho é necessário definir o conceito de séries temporais, que é um conjunto de m variáveis dadas por T= t1, t2, …, tm que normalmente são organizados por ordem temporal e espaçadas em intervalos de tempo iguais [3]. Para identificar os dados desejados precisamos buscar os padrões dessas curvas e a partir desses padrões reconhecer curvas de exoplanetas para qualquer base de dados. Para cumprir com esse objetivo utilizamos algoritmos de aprendizado de máquina, os quais baseiam-se na construção de um modelo de inferência a partir de uma base de dados conhecida, de forma a possibilitar a classificação de novos dados (desconhecidos) automaticamente, sem a interferência de especialistas. Neste trabalho utilizamos esses algoritmos em combinação com os dados gerados pelo processamento de recorrências. Esse processamento consiste da utilização de um gráfico projetado para localizar padrões recorrentes que sejam

aparentemente ocultos para o observador e pode ser explicado a partir do teorema de Takens. Pelo teorema, podemos recriar uma imagem topologicamente equivalente do comportamento do sistema multidimensional original usando a série temporal de uma única variável observável, ou seja para a série x_i, construímos vetores do tipo $x_i^m = (x_i, x_{i+d}, x_{i+2d}, \ldots, x_{i+(m-1)d})$, onde m é a dimensão de incorporação e d é o atraso de tempo. Em seguida, uma matriz simétrica de distâncias pode ser construída calculando distâncias entre todos os pares de vetores embutidos; o gráfico de recorrência relaciona cada distância de tal matriz a uma cor, assim o gráfico de recorrência é um gráfico retangular sólido que consiste em pixels cujas cores correspondem à magnitude dos valores dos dados em uma matriz bidimensional e cujas coordenadas correspondem às localizações dos valores de dados na matriz [1].

3 METODOLOGIA

Neste trabalho analisamos as curvas de luz provenientes da base de dados do Kaggle (www.kaggle.com), a qual é composta por um subconjunto de dados extraídos do projeto Kepler. Nesses dados, cada estrela tem um rótulo binário de 2 ou 1, o 2 indica que a estrela está confirmada para ter pelo menos um exoplaneta em órbita. Nosso conjunto de dados é composto por 42 estrelas confirmadas com exoplanetas e 5615 estrelas sem exoplanetas, sendo que cada uma destas curvas possui um total de 3197 registros. A abordagem utilizada foi tomar estes dados originais e aplicar a função de análise de recorrência. Esse processo nos fornece uma matriz de recorrência para cada curva de luz, as quais podemos visualizar por meio dos exemplos representados na Figura 1.

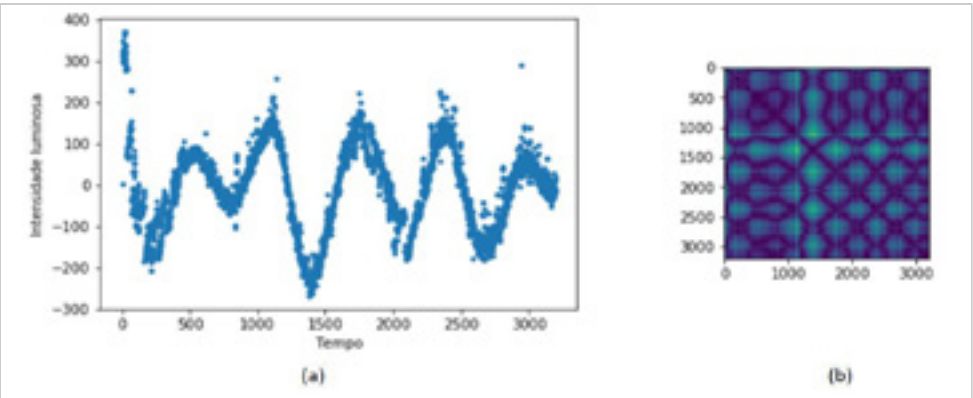

Figura 1: (a) exemplo de curva de luz; (b) gráfico de recorrência gerado.

O processo descrito acima foi realizado mediante a utilização da biblioteca sklearn (Scikit-Learn para Python), a qual reúne vários métodos de algoritmos de classificação, regressão e agrupamento de dados. Também, desenvolvemos um programa computacional para aplicar os algoritmos de aprendizagem de máquina, nos quais estão incluídos algoritmos como o Decision Trees, Support Vector Machines, Naive Bayes, Nearest Neighbors e Neural Network. A análise do desempenho dos algoritmos de classificação foi realizado por meio da biblioteca de validação cruzada. Como medida de desempenho utilizamos o escore F1, que pode ser interpretado como uma média ponderada da *precision* e da *recall* (métricas de precisão), em que um escore F1 alcança seu melhor valor em 1 e o pior escore em 0. A equação para a pontuação de F1 é:

$$F1 = 2 * (precision * recall) / (precision + recall) \quad (1)$$

Para a execução desses códigos contamos com a utilização do Cluster C3HPC (DINF-UFPR) que possui 6 nodos de processamento, cada um com 4 sockets 3.30GHz (8 núcleos por socket) e 256 GB de RAM.

4 RESULTADOS E DISCUSSÃO

Executamos dois experimentos com os algoritmos de aprendizado de máquina: (1) para os dados originais das séries temporais e (2) para as matrizes obtidas a partir das análises de recorrência. Do conjunto de dados do Kepler, utilizamos as 42 curvas de luz de estrelas com exoplanetas e apenas 158 curvas de não exoplanetas, devido a limitação computacional. Os resultados dos experimentos estão apresentados na Tabela 1.

Algoritmos	Médias dados originais (1)	Médias análise de recorrência(2)
decision tree	81,78% ± 1,85%	91,67% ± 2,32%
vector machine	88,27% ± 0,50%	89,02% ± 0,89%
nearest neighbors	89,73% ± 1,58%	89,67% ± 3,60%
naive bayes	72,37% ± 27,93%	98,14% ± 1,51%
neural networks	72,26% ± 8,30%	88,29% ± 1,23%

Tabela 1: Resultados com F1 score.

Observamos melhores resultados nos dados classificados com a aplicação da análise de recorrência, com exceção do algoritmo nearest neighbors, para o qual em

ambos os casos apresentou médias muito similares. Por outro lado, podemos observar significativa melhora nos resultados obtidos com os algoritmos de decision tree e naive bayes.

5 CONCLUSÕES

Com base nos resultados apresentados neste trabalho percebemos em alguns algoritmos uma proporção maior de acertos para os dados com análise de recorrência, o que nos permite concluir pela análise da métrica F1 que essa técnica possibilitou melhoria no processo de classificação. Porém, acreditamos que estes resultados poderiam ser ainda melhores, o que não foi possível por motivos como a baixa amostragem de objetos que eram estrelas com exoplanetas. Logo, em projetos futuros pretendemos utilizar bases de dados com uma menor divergência na quantidade de itens e ainda dados em que tenhamos a certeza da existência de apenas duas classes, que é o tipo de análise que determinamos para os algoritmos de aprendizado de máquina que utilizamos.

6 PRINCIPAIS REFERÊNCIAS BIBLIOGRÁFICAS

[1] BELAIRE-FRANCH, Jorge. CONTRERAS, Dulce. Recurrence Plots in Nonlinear Time Series Analysis: Free Software. Dept. of Economic Analysis University of Valencia, 2002.
[2] MITCHELL, T. M. Machine Learning. Boston, USA: McGraw-Hill, 1997.
[3] ZALEWSKI, Willian. Modelagem Simbólica de Padrões Morfológicos para a Classificação de Séries Temporais. Curitiba, PR, p. 55-58, 2015.

7 AGRADECIMENTOS

Agradecimento a UNILA pelo financiamento da bolsa ITI - Iniciação Tecnológica que contribuiu para o desenvolvimento deste trabalho.