ESTUDO E IMPLEMENTAÇÃO DE ALGORITMOS DE DETECÇÃO E CLASSIFICAÇÃO EM REDES COMPLEXAS EMPREGANDO A TEORIA DOS GRAFOS

Samuel Antunes Miranda [1]; Vitor Alex Oliveira Alves [2]

[1] Aluno de Iniciação Científica da Escola de Engenharia Mauá (EEM/CEUN-IMT);
[2] Professor da Escola de Engenharia Mauá (EEM/CEUN-IMT).

Resumo. *O artigo descreve um sistema de processamento de séries temporais, no presente caso curvas de luz oriundas do telescópio espacial CoRot, com o objetivo de classificar tais séries como representativas da observação de exoplanetas ou de sistemas estelares binários. As séries temporais são pré-processadas para homogeneizar seu comprimento e período de amostragem e posteriormente são transformadas em redes complexas (grafos) e então convertidas em gráficos de recorrência. Tais gráficos são as informações de entrada de uma rede neural convolucional que procura classifica-las como geradas pela observação de um exoplanetas ou de um sistema de estrelas eclipsantes binárias. Os resultados obtidos com o uso dessa abordagem são promissores, atingindo índices de acerto nos dados de validação do sistema superiores à 84%.*

Introdução

Com o advento da Inteligência Artificial, uma enorme quantidade de informação vem sendo coletada a partir dos processos do mundo atual. Essas informações, cada vez mais, agregam em si novas formas de complexidade (Mayer-Schönberger e Cukier, 2013). Este fato impõe grandes desafios aos pesquisadores de diversas áreas, na medida em que se torna necessário trabalhar em conjunto para se extrair padrões ou novas estruturas de interação sistêmica a partir de dados de grande volume, com grande diversidade de informações e coletados em altíssima velocidade. Técnicas avançadas de análise de dados, tais como *Machine Learning* e *Data Mining*, procuram detectar estruturas ou padrões escondidos em meio ao ambiente caótico das informações coletadas. Assim, é possível transformar dados com correlações ou interações aparentemente incompreensíveis à primeira vista em um conjunto restrito de parâmetros confiáveis a partir do qual seja viável uma tomada de decisão assertiva (Zou *et al.*, 2019).

Em caráter complementar a tais técnicas, há grande espaço para o desenvolvimento de novas ferramentas para a análise de dados aplicáveis a redes complexas (Albert e Barabasi, 2002). Estes construtos são capazes de modelar as relações entre as diversas entidades que formam o universo em estudo (conjunto de dados). A título de exemplo, as entidades podem ser fontes de emissão de ondas de rádio, veículos detectados em câmeras de vigilância, pessoas mencionadas em documentos sensíveis à segurança nacional de determinado país, operações e transações bancárias (Miller *et al.*, 2013) ou ainda, séries temporais geradas a partir de curvas de luz estelar (foco de aplicação deste projeto de iniciação científica). A análise das relações entre as entidades pode proporcionar melhoria significativa na capacidade dos analistas de encontrar interações sutis, planejadas ou não, que passariam despercebidas em meio a milhões de outras interações se o componente relacional dos dados não fosse considerado na análise. Alguns exemplos de questões que podem ser respondidas a partir de análises de interações: quais emissores de ondas de rádio estão presentes em uma mesma localidade? Quais veículos estacionaram em uma mesma área monitorada? Quais pessoas são mencionadas em um mesmo documento? Quais transações bancárias são fraudulentas? Quais estrelas de um mesmo aglomerado compartilham determinada característica? Como detectar exoplanetas a partir de curvas de luz?

Recentemente, a comunidade científica tem devotado grande esforço na aplicação da teoria de redes complexas no contexto da análise de séries temporais (Kantz e Schreiber, 2004; Sprott, 2003). Séries temporais são sequências de dados indexados de forma ordenada ao longo do tempo e sua análise considera o estudo de todas as observações ao invés de instâncias individuais. Com o intuito de se descobrir padrões ocultos em grandes conjuntos de dados de diferentes fontes,

ferramentas de Data Mining desenvolvidas em pesquisas no campo da Ciência da Computação passaram a ser aplicadas a séries temporais com propósitos de *clustering*, classificação e previsão de eventos (Keogh e Kasetty, 2003; Aghabozorgi *et al.*, 2015), em diversas áreas como meteorologia, economia, medicina e astronomia.

Em conclusão, a teoria de detecção de anomalias e classificação em redes complexas fundamentadas em dados relacionais é uma importante área de pesquisa nos dias atuais. Consequentemente, o estudo e a implementação de algoritmos capazes de realizar as tarefas previamente mencionadas constituem trabalho altamente relevante no cenário atual. Este projeto de iniciação científica objetiva o estudo e a implementação de um algoritmo de modelagem de séries temporais sob a forma de redes complexas, assim como o estudo e a implementação de um algoritmo de detecção e classificação de anomalias em tais redes. Como plataforma de testes, os algoritmos desenvolvidos foram aplicados no estudo de curvas de luz estelar (série temporal que modela a variação do brilho de uma estrela ao longo do tempo) geradas pelo telescópio espacial CoRot. O projeto CoRoT (acrônimo para COnvecção + ROtação + Trânsitos) foi desenvolvido pela Agência Espacial Francesa (CNES) em conjunto com vários laboratórios franceses e parceiros internacionais, incluindo o Brasil. O Instituto Mauá de Tecnologia participou do esforço, com o Prof. Dr. Vanderlei Cunha Parro trabalhando como pesquisador visitante no Observatório de Paris no período de 2004 a 2006. A avaliação dos dados fornecidos pelo telescópio se prestará à detecção de anomalias (eclipses) e possível classificação de tais eventos em função do objeto causador (discernimento entre eclipses gerados por exoplanetas ou diferentes tipos de estrela).

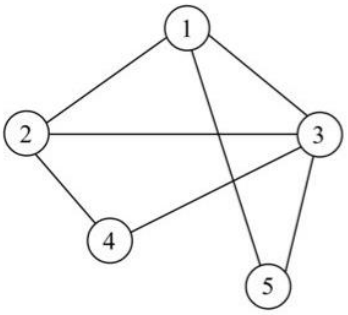
Algumas definições importantes

As curvas de luz empregadas neste estudo constituem séries temporais de comprimentos variáveis (essencialmente, o número de pontos da série temporal depende do intervalo de amostragem e do tempo de captura) em que o valor observado é a intensidade luminosa detectada pelo receptor quando apontado para determinada região do universo. Detecções de variações bruscas na intensidade luminosa são indícios de elipses, que podem ser causados por exoplanetas ou por estrelas binárias. O propósito deste estudo é justamente classificar as curvas de luz disponíveis como sendo representativas destes dois tipos de corpos celestes.

O primeiro passo de análise é representar as séries temporais no formato de redes complexas, para a posterior geração dos gráficos de recorrência (RP, do inglês *recurrence plots*). São estas representações gráficas que serão analisadas por redes neurais convolucionais e então classificadas como eclipses causados por exoplanetas ou por estrelas binárias.

Definição 1. Uma rede complexa é frequentemente representada como um grafo G=(V,E) que consiste de dois conjuntos V e E. O conjunto V é a coleção de vértices (nós, pontos) de G, enquanto E é o conjunto de arestas (ligações, linhas) representando pares de elementos conectados de V (Costa *et al.*, 2007). Cada vértice é identificado por um índice inteiro p=1,...,N e cada aresta é identificada por um par (p,q), identificando uma ligação entre os vértices p e q. Um grafo G é chamado não-direcionado se uma aresta unindo os vértices p e q, denotada por (p,q), é equivalente à aresta (q,p) que liga os vértices q e p, ou seja, $(p,q)\in E \Leftrightarrow (q,p)\in E$. Por ouro lado, em um grafo direcionado, tipicamente temos $(p,q) \in E \not\Leftrightarrow (q,p) \in E$. Um grafo pode conter laços (ou malhas), ou seja, arestas que ligam um vértice a si mesmo. De forma geral, podem ser atribuídos pesos W_{pq} às arestas (p,q). Neste caso, um grafo direcional com pesos pode ser completamente descrito por sua matriz de pesos **W**, na qual cada entrada expressa o peso da conexão do vértice p para o vértice q.

Definição 2. A matriz (binária) de adjacência **A** de um grafo sem pesos é tal que $A_{pq}=1$ indica a presença de uma aresta (p,q) no conjunto E; caso não haja conexão entre os vértice p e q, tem-se $A_{pq}=0$. Dependendo da estratégia empregada na transformação de uma série temporal em uma rede complexa, a resultante matriz de adjacência **A** apresentará dependência sobre os parâmetros do algoritmo empregado. Neste estudo, tal dependência se dá com relação ao parâmetro ε da rede de recorrência. A Figura 1 mostra um grafo de 5 vértices e 7 arestas, em conjunto com sua respectiva matriz de adjacência.



	V1	V2	V3	V4	V5
V1	0	1	1	0	1
V2	1	0	1	1	0
V3	1	1	0	1	1
V4	0	1	1	0	0
V5	1	0	1	0	0

Figura 1 – Exemplo de um grafo e sua respectiva matriz de adjacência

Definição 3. A transformação de uma série temporal para o domínio das redes complexas se dá pela análise da similaridade estatística ou de uma métrica de proximidade entre os diferentes segmentos da série temporal. Neste estudo, são aplicadas as redes de recorrência (Donner *et al.*, 2010; Marwan *et al.*, 2009) que propiciam uma reinterpretação dos gráficos de recorrência. O leitor interessado nos detalhes técnicos é referenciado à (Marwan *et al.*, 2007). Em linhas gerais, interpreta-se uma série temporal $\{x_i\}_{i=1}^{N}$, com $x_i=x(t_i)$, como a representação finita de uma trajetória descrita por um sistema dinâmico. A seguir, os dados são convertidos em vetores de estado com dimensão apropriada. Nas redes de recorrência, os estados tornam-se arbitrariamente próximos dos anteriores após certo tempo, sendo esta uma propriedade fundamental de sistemas dinâmicos determinísticos e também é típica de sistemas não-lineares caóticos (Ott, 2002). Essas recorrências podem ser visualizadas por meio dos gráficos de recorrência (Eckmann *et al.*, 1987), que constituem a representação gráfica da matriz de recorrência da rede $\mathbf{R}(\varepsilon)$ definida por:

$$R_{ij}(\varepsilon) = \Theta(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (1)$$

em que ||.|| pode ser a norma Euclidiana, Norma Máxima, Manhattan, etc e Θ é uma função do tipo heaviside. Os gráficos de recorrência permitem investigar a recorrência de uma trajetória em um espaço de fase m-dimensional por meio de uma representação bidimensional de Rij em termos de pontos que indicam pares recorrentes e não recorrentes de vetores de estado. O parâmetro ε é um limitante pré-definido que determina se dois vetores de estado são próximos ou não. O passo crucial é reinterpretar $\mathbf{R}(\varepsilon)$ como a matriz de adjacência **A** de uma rede complexa adjunta embutida no espaço de fase,

$$\mathbf{A}(\varepsilon) = \mathbf{R}(\varepsilon) - \mathbf{I}_N, \quad (2)$$

Sendo assim, os vetores de estado **x**i são interpretados como vértices de uma rede complexa que são conectados por arestas não direcionadas se são mutualmente próximos no espaço de fases, ou seja, caso representem recorrências.

Definição 4. As Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Network*) formam uma classe de rede neural artificial do tipo *feed-forward* com aplicação no processamento e análise de imagens digitais. As CNN empregam uma variação da estrutura de *perceptrons* multicamada com o objetivo de minimizar o pré-processamento. O padrão de conectividade das CNN é inspirado na organização do córtex visual dos animais, razão pela qual essa classe de redes neurais é amplamente empregada na análise de imagens. Tratamentos detalhados sobre o assunto são vistos em (Michelucci, 2019; Kelleher, 2019).

Material e Métodos

As séries temporais (curvas de luz) disponíveis para análise possuem diferentes comprimentos e também diferentes períodos de amostragem. Assim, um procedimento de pré-processamento de dados foi necessário para ajustar o comprimento (número de pontos), o chamado *resize*, e ajustar o período de amostragem, o chamado *resample*. De posse das curvas de luz pré-processadas, passa-se à etapa de conversão dessas séries temporais em redes de recorrência e a

posterior transformação dessas entidades em gráficos de recorrência. Estas representações gráficas são ilustrações de matrizes NxN oriundas de séries temporais com N estados e apresentam simetria com relação à diagonal principal.

Uma porção dos gráficos de recorrência (80%) é destinada ao processo de treinamento da rede neural convolucional. Durante a fase de treinamento, é computada a função perda, um índice de mérito que avalia a quantidade de classificações corretas dentro do universo de informações investigado. Um treinamento bem-sucedido implica na minimização da função perda. Em um procedimento de validação cruzada, os outros 20% dos gráficos de recorrência são empregados como dados de teste para comprovar a eficácia da classificação fornecida pela rede neural, no pressente caso discernindo entre exoplanetas e eclipsantes binárias.

A linguagem de programação empregada foi o Python, em sua versão 3.8.3, assim como as bibliotecas *numpy* 1.19.2, *skimage* 0.17.2, *matplotli*b 3.2.2 e *keras* 2.4.3. As rotinas de conversão das curvas de luz em gráficos de recorrência empregaram a biblioteca *pycorn*, desenvolvida por Zou *et al.* (2019). A configuração da rede neural convolucional, os processos de treinamento e teste também foram desenvolvidos em Python. A Figura 2 ilustra o processo descrito.

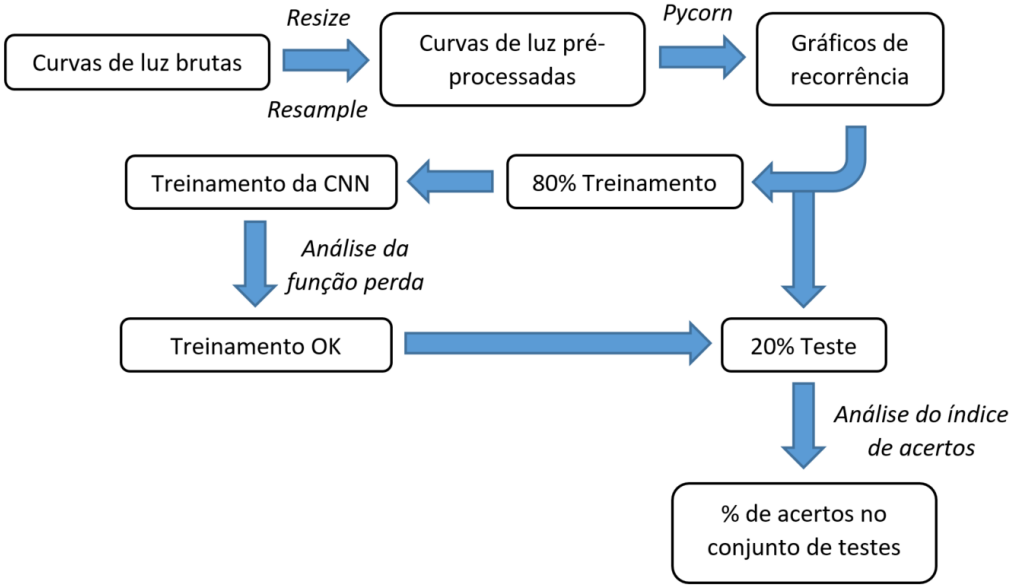


Figura 2 – Etapas de processamento para a classificação das curvas de luz.

Como mencionado anteriormente as CNN são amplamente utilizadas em aplicações que lidam com imagens. De fato, a técnica se fundamenta na inserção da imagem em filtros que procuram, a cada iteração do processo de treinamento, aprimorar a detecção de características que permitam discernir entre as classificações possíveis. A Tabela 1 descreve a arquitetura específica utilizada neste trabalho.

Tabela 1 – Arquitetura da CNN

Tipo de camada	Formato de entrada	Formato de saída
1ª Convolucional 2D	128 x 128 x 8	126 x 126 x 8
1ª *MaxPooling*	126 x 126 x8	63 x 63 x 8
1ª *Dropout*	63 x 63 x 8	63 x 63 x 8
2ª Convolucional 2D	63 x 63 x 8	61 x 61 x 8
2ª *MaxPooling*	61 x 61 x 8	30 x 30 x 8
2ª *Dropout*	30 x 30 x 8	30 x 30 x 8
Linearização	30 x 30 x 8	7200
1ª Densa	7200	32
3ª *Dropout*	32	32
2ª Densa	32	2

A Figura 3 fornece uma visualização da arquitetura proposta para a CNN. A ordem das camadas é a mesma fornecida na Tabela 1.

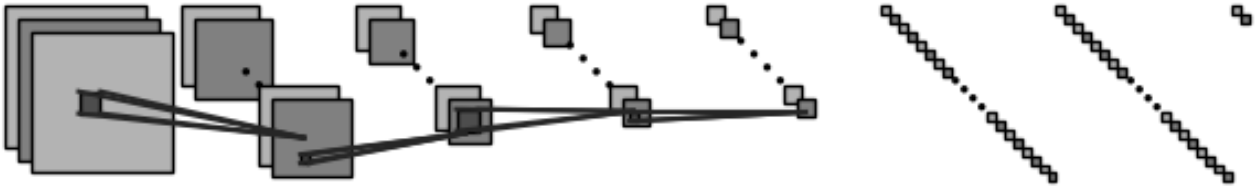


Figura 3 – Visualização da arquitetura da CNN.

Resultados e Discussão

As Figuras 4 e 5 ilustram alguns exemplos de curvas de luz convertidas em gráficos de recorrência (RP). É importante notar nos RP's a simetria com relação a diagonal principal e os padrões de imagem formados em ambos os tipos de fenômenos observados, exoplanetas (Figura 4) e estrelas eclipsantes binárias (Figura 5).

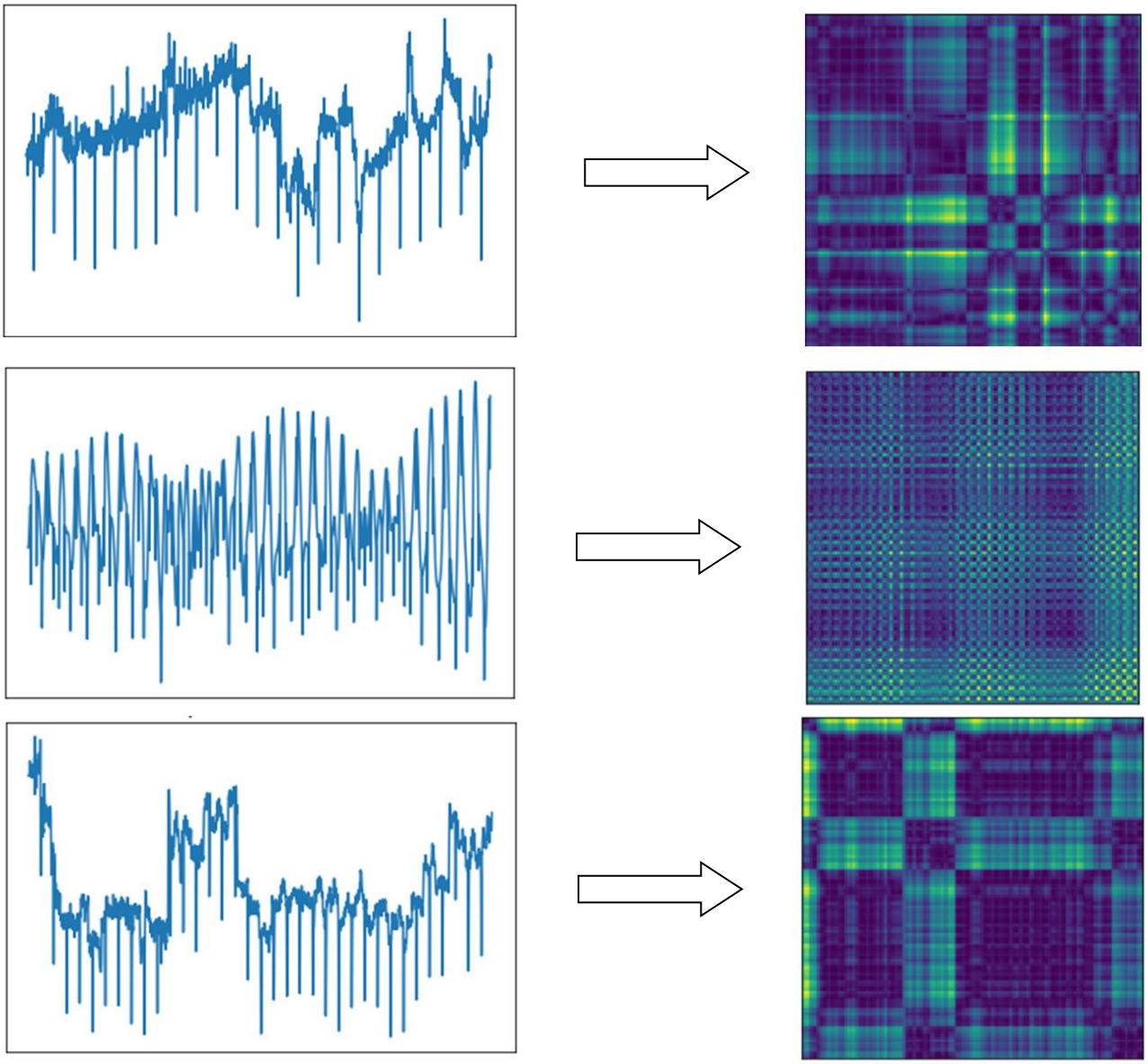


Figura 4 – Curvas de luz nas quais foram observados exoplanetas e suas respectivas representações no formato de gráficos de recorrência.

O processo de treinamento da CNN não deve ser único. De fato, é aconselhável repeti-lo diversas vezes, empregando em cada etapa diferentes conjuntos de treinamento e teste. Neste estudo, 100 treinamentos foram realizados e as respectivas taxas de acerto no treinamento e na validação (teste com os 20% de dados não empregados no treinamento) foram calculadas. Os três melhores resultados estão registrados na Tabela 2.

Tabela 2 – Três melhores índices de acerto dentre 100 treinamentos (ordenados pelo índice obtido na etapa de validação)

Índice de acerto nos dados de treinamento	Índice de acerto nos dados de validação
84,2%	86,6%
92,3%	85,3%
86%	84,7%

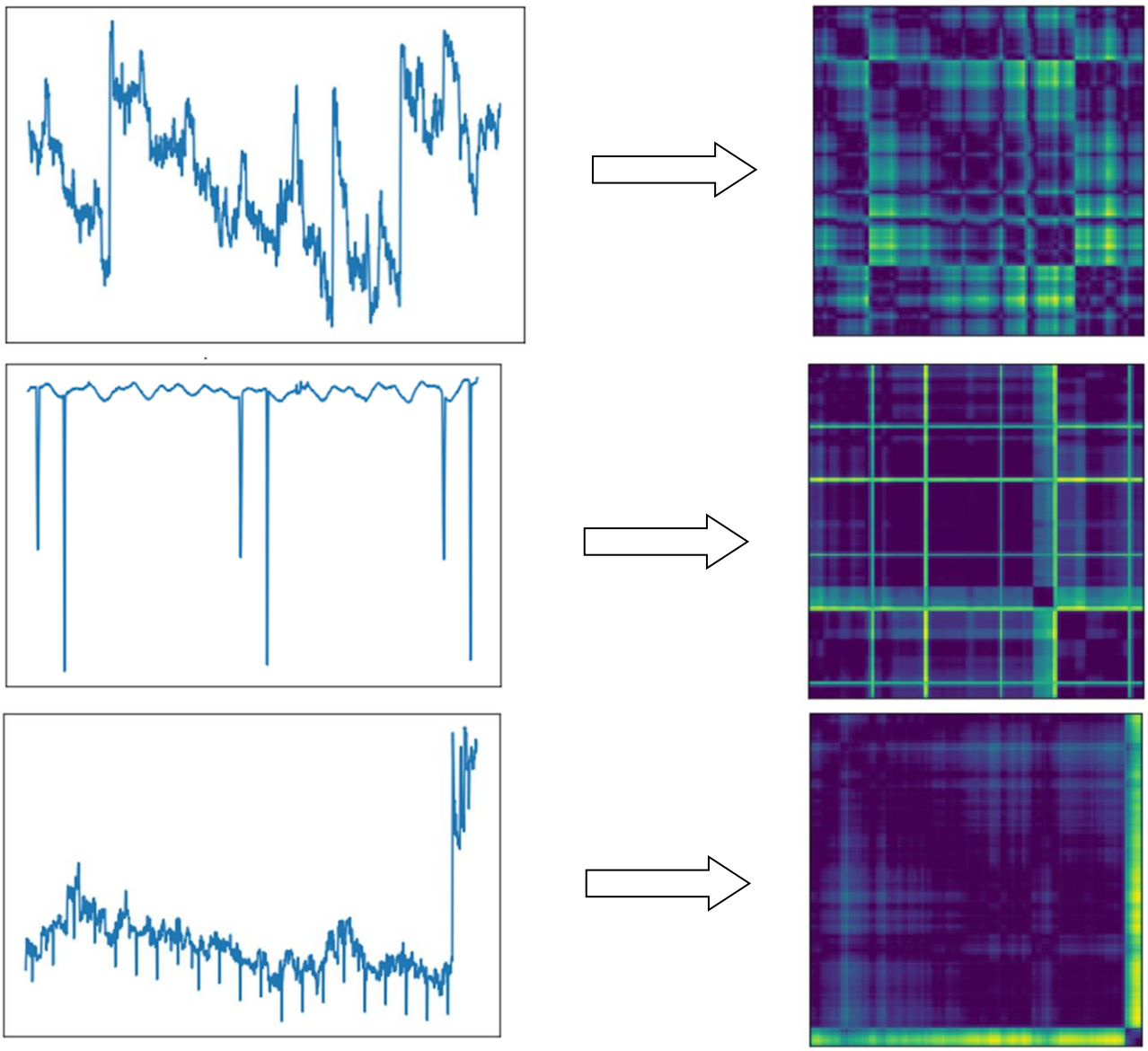
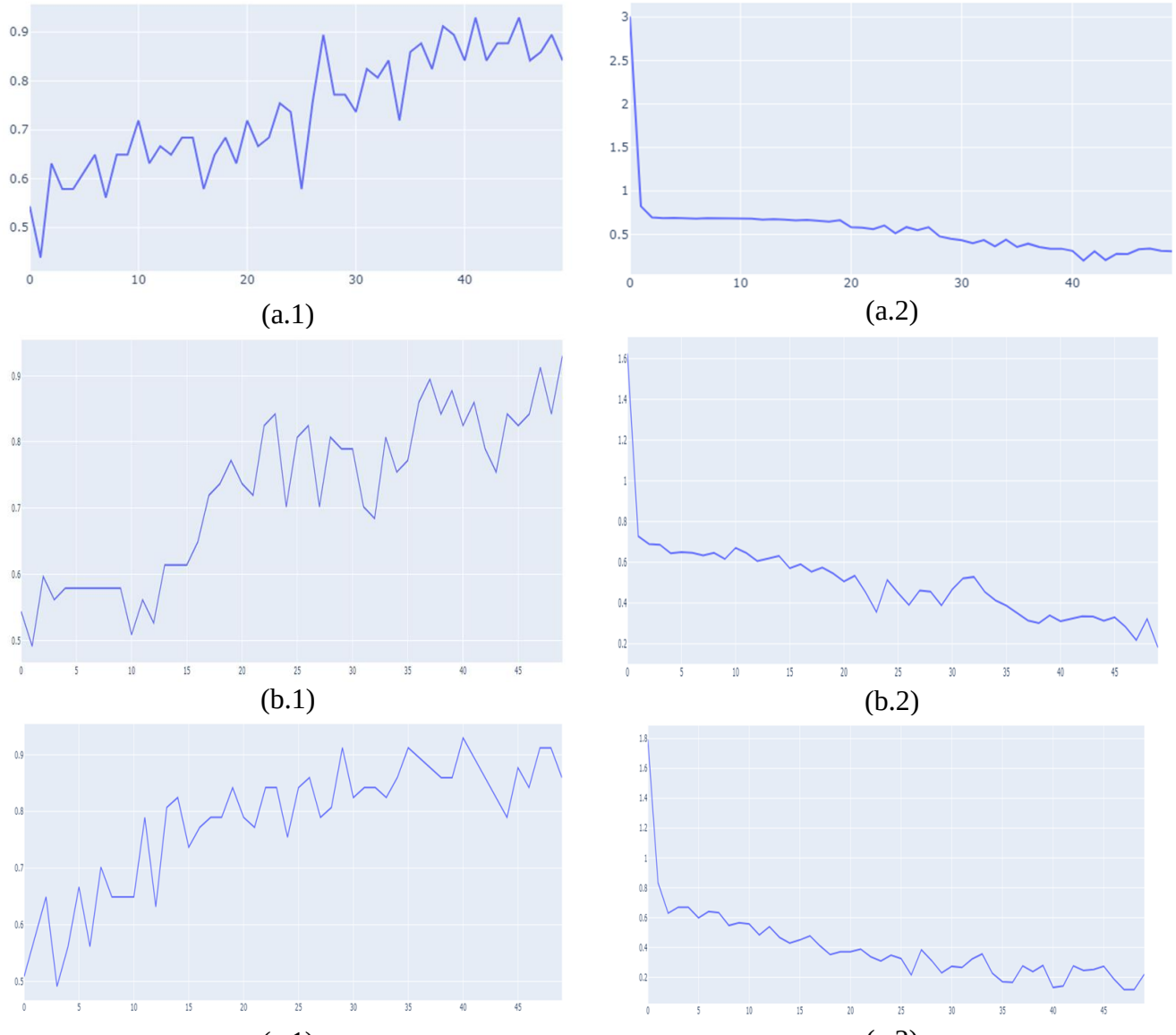


Figura 5 – Curvas de luz nas quais foram observadas estrelas eclipsantes binárias e suas respectivas representações no formato de gráficos de recorrência.

As informações da Tabela 2 revelam que nem sempre um alto índice de acertos na etapa de treinamento implica em melhor eficácia na etapa de validação. Esta é uma característica intrínseca

deste tipo de abordagem que emprega redes neurais convolucionais como classificadores. Isto evidencia a importância de se trabalhar exaustivamente com os dados disponíveis e também promover modificações na arquitetura da CNN com o objetivo de maximizar o índice de acertos nos dados de validação. De forma contra intuitiva, observa-se que o melhor índice de certos nos dados de validação ocorreu para a rede que exibiu (dentre os três melhores resultados) o pior índice de acertos nos dados de treinamento. Novamente, o que se procura é uma solução de compromisso entre complexidade da rede (o que impacta em tempo de treinamento) e um índice de erros abaixo de um valor considerado aceitável.

A Figura 6 mostra a representação gráfica dos índices de acerto (à esquerda) e das funções perda (à direita), durante a fase de treinamento da CNN, para as realizações registradas na Tabela 2. É possível notar a tendência de crescimento dos gráficos relativos aos índices de acerto, o que revela o aumento da capacidade do modelo de discernir entre as duas classificações possíveis, a cada iteração do processo. Concomitantemente, nota-se a diminuição das funções perda associadas à cada CNN.



(a.1) (a.2)

(b.1) (b.2)

(c.1) (c.2)

Figura 6 – Representações das evoluções dos índices de acerto (x.1) e das respectivas funções perda (x.2) obtidas pelas CNN da Tabela 2. As letras a, b e c representam, respectivamente, as linhas da Tabela 2.

Conclusões

Uma abordagem de classificação empregando CNN foi descrita neste estudo. O objetivo é discernir quais curvas de luz são representativas das observações (feitas pelo CoRoT) de exoplanetas ou de estrelas binárias eclipsantes. Para tal, um sistema de processamento foi proposto (vide Figura 2), no qual as curvas de luz são pré-processadas e posteriormente transformadas em gráficos de recorrência que constituem os dados de entrada da CNN classificadora.

Os resultados obtidos revelam que tal abordagem é promissora, fato evidenciado pelos índices de acerto em dados de validação superiores à 80%. Nos três melhores resultados, dentre os 100 treinamentos realizados, tais índices alcançaram valores de 86,6%, 85,3% e 84,7%.

Como possíveis passos futuros relativos a este projeto, podem ser citados (i) testes do sistema em larga escala (com o uso de um banco de dados maior), seja com relação ao número de amostras em cada série temporal, seja com relação ao número de séries temporais disponíveis ou mesmo com a expansão do número de categorias; (ii) a inserção de outros algoritmos de detecção de anomalias em redes complexas; (iii) a inclusão de outras arquiteturas de redes neurais.

Referências Bibliográficas

Aghabozorgi, S.; Shirkhorshidi, A.S.; Ying Wah, T. (2015) Time-series clustering - A decade review, *Inf. Syst.* **53** (2015) 16–38.

Albert, R.; Barabasi A. L. (2002) Statistical mechanics of complex networks, *Rev. Modern Phys.* **74** (1) 47–97

Costa, L.F.; Rodrigues, F.A.; Travieso, G.; Villas Boas, P.R. (2007) Characterization of complex networks: A survey of measurements, *Adv. Phys.* **56** (1)167–242.

Donner, R.V. ; Zou, Y.; Donges, J.F.; Marwan, N.; Kurths J. (2010) Recurrence networks — a novel paradigm for nonlinear time series analysis, *New J. Phys.* **12** (3) 033025.

Eckmann, J.P.; Kamphorst, S.O.; Ruelle D. (1987) Recurrence plots of dynamical systems, *Europhys. Lett.* **4** (9) 973–977.

Kantz H.; Schreiber, T. (2004) *Nonlinear Time Series Analysis*, 2nd edition. Cambridge University Press.

Kelleher, J. D. (2019) *Deep Learning.* The MIT Press.

Keogh, E.; Kasetty, S. (2003) On the need for time series data mining benchmarks: A survey and empirical demonstration, *Data Min. Knowl. Discov.* **7** (4) 349–371.

Marwan, N.; Romano, M.C.; Thiel, M.; Kurths J. (2007) Recurrence plots for the analysis of complex systems, *Phys. Rep.* **438** (5–6) 237–329.

Marwan, N.; Donges, J.F.; Zou, Y.; Donner, R.V.; Kurths J. (2009) Complex network approach for recurrence analysis of time series, *Phys. Lett.* A **373** (46) 4246–4254.

Mayer-Schönberger, V.; Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

Michelucci, U. (2019) *Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection.* Apress.

Miller, B.A.; Bliss, N.T.; Wolfe, P.J.; Beard, M.S (2013) Detection Theory for Graphs, *Lincoln Laboratory Journal* **20** (1) 10-30.

Ott, E (2002) *Chaos in Dynamical Systems.* 2nd edition. Cambridge, Cambridge University Press.

Sprott, J.C. (2003) *Chaos and Time-Series Analysis*, Oxford University Press, Oxford.

Zou Y.; Donner, R. V.; Marwan, N.; Donges, J. F.; Kurths J. (2019) Complex Network approaches to nonlinear time series analysis. *Phys. Rep.* **787** 1–97.