Anais do V Congresso Brasileiro de Eletromiografia e Cinesiologia e X Simpósio de Engenharia Biomédica - ISBN: 978-85-5722-065-2 - DOI: 10.29327/cobecseb.78820

# MINERAÇÃO DE DADOS APLICADA À SÉRIES TEMPORAIS RESULTANTES DE SINAIS MOTORES DA DOENÇA DE PARKINSON

J. P. Folador* e A. O. Andrade*

*Núcleo de Inovação e Avaliação Tecnológica em Saúde (NIATS), Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, Brasil
e-mail: jpfolador@gmail.com

**Resumo:** Nos últimos anos, a quantidade de dados gerados na área da saúde tem aumentado consideravelmente. Neste sentido, é necessário extrair informações destes dados, a fim de auxiliar na tomada de decisão. O trabalho aqui proposto consiste na revisão do estado atual da literatura e identificação de técnicas de mineração de dados que têm sido utilizadas na extração de informações de séries temporais, e que resultem do registro de sinais na área da engenharia biomédica e sinais motores da doença de Parkinson. Os resultados da revisão sugerem que os métodos mais utilizados na mineração de dados representados por séries temporais são: Árvore de Decisão, Redes Neurais, Redes Bayesianas e Modelos Ocultos de Markov. Ademais, observou-se que existe uma necessidade da popularização do uso destas técnicas na avaliação de séries temporais representando os sinais motores da Doença de Parkinson, em especial o surgimento de aplicações relacionados ao monitoramento prolongado dos sinais motores da doença, e consequente o aumento no volume de dados a serem processados.
**Palavras-chave:** séries temporais, mineração de dados, doença de Parkinson.

***Abstract:*** *In recent years, the amount of generated information in the health sector has considerably increased. In this way, it is necessary to extract information from these data, in order to help to take decisions on the data interpretation. The work proposed here consists in the review of the current literature to identify the most used data mining techniques applied to time series in Biomedical engineering area and motor signs of Parkinson's disease. The review suggests the most used methods for data mining represented by time series are Decision tree, neural networks, Bayesian networks and Hidden Markov Models. Moreover, we observed that there is a necessity of popularization of techniques used in time series analysis representing the motor signs of Parkinson's disease, especially when we consider the increase of related applications to monitoring prolonged activities of the disease and consequent increase in the volume of data to be processed.*
***Keywords:*** *time series, data mining, Parkinson's disease.*

## Introdução

Uma série temporal pode ser gerada em diferentes domínios, contendo um grande volume de dados e pode ser heterogênea. De forma simples, uma série temporal é gerada a partir de observações cronológicas com medidas sucessivas de certas quantidades. Este formato de dado coletado no tempo é bastante usado em inúmeros aplicações e diversas áreas do conhecimento [16]. Com cerca de 40% da população mundial conectada à internet [11] o volume de dados gerados hoje, chega a 2,7 Zettabytes (2,7 $10^{12}$ GB) e, até 2020, pode chegar a 35 Zettabytes [21]. Na área da saúde em 2012, tinha-se cerca de 500 Petabytes, e até 2020 espera-se um aumento de 50 vezes no volume de informações geradas [18].

Neste crescente volume de dados sendo coletados na área da saúde, os estudos dos sinais cardinais da doença de Parkinson (DP) também tem aumentado em relação ao monitoramento de longo prazo (e.g., monitorado por horas ou dias) [1]. Assim, este trabalho visa conceituar mineração de dados e séries temporais, apontar as técnicas mais usadas na extração de informações de sinais coletados na área da engenharia biomédica e verificar se estas técnicas tem sido relacionadas com a DP.

## Materiais e Métodos

A revisão sistemática da literatura possibilita compreender os conceitos básicos, o estado da arte da temática levantada e encontrar possíveis áreas de estudo. Neste sentido, objetivou responder as seguintes questões:

- Como é feita a mineração em séries temporais?
- Quais as técnicas mais utilizadas na mineração de dados representados por séries temporais resultantes de sistemas comumente estudados na Engenharia Biomédica?
- Existem trabalhos sobre mineração de dados em sinais cardinais da doença de Parkinson?

Para alcançar os objetivos foram realizadas consultas exploratórias em pelo menos quatro diferentes bases de dados de periódicos, a destacar: PubMed (predomínio de literatura biomédica), IEEE *Xplorer* (literatura voltada para engenharia e computação), *Scopus* (ciências, tecnologia e medicina) e *Science Direct* (ciências, saúde e tecnologia). No total, 5 *strings* de busca foram criadas, com o intuito de se realizar consultas nas bases de dados:

- ***String 1*****:** "*time series*" AND "*data mining*";
- ***String 2*****:** "*time series*" AND "*data mining*" AND "*biomedical engineering*";
- ***String 3:*** "*time series*" AND "*data mining*" AND "*Parkinson's disease*";
- ***String 4:*** "*data mining*" AND "*Biomedical engineering*" AND "*Parkinson's disease*";

- ***String 5:*** "*time series*" AND "*data mining*" AND "*Biomedical engineering*" AND "*Parkinson's disease*"

As *strings* de pesquisa refletem a busca de informações para responder os objetivos. E, para inclusão ou exclusão dos trabalhos nessa revisão, foram analisados os resumos, de modo a verificar a relevância dos mesmos, e os trabalhos encontrados nos últimos 10 anos visou buscar técnicas de mineração de dados atuais sobre a engenharia biomédica e a DP. Foram consideradas as *strings* de busca presentes no título, no resumo ou nas palavras-chave. Somente para a *string* 1 foram considerados trabalhos de revisão, pois objetivou-se, nesta etapa, a busca conceitual de série temporal e mineração de dados.

**Resultados**

A quantidade de trabalhos encontrados relacionados a cada *string* de busca, é mostrada na Tabela 1. As buscas realizadas com a *String* 1 resultaram no maior número de estudos, visto que a mesma aborda conceitos genéricos sobre séries temporais e mineração de dados.

**Tabela 1**: Artigos encontrados em cada base de dados.

| String | IEEE Xplore | | Scopus | | Science Direct | | PubMed | |
|---|---|---|---|---|---|---|---|---|
| | E | A | E | A | E | A | E | A |
| String 1 | 28 | 6 | 26 | 4 | 10 | 0 | 6 | 1 |
| String 2 | 21 | 3 | 8 | 3 | 0 | 0 | 3 | 1 |
| String 3 | 4 | 2 | 6 | 1 | 1 | 1 | 2 | 0 |
| String 4 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 0 |
| String 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Total** | **58** | **14** | **42** | **8** | **11** | **1** | **11** | **2** |

**Legenda**: E – Encontrados, A - Aproveitados

Após a leitura do resumo de cada trabalho foram selecionados 25 artigos, ou seja, 20,5% do total de 122 artigos; os demais foram excluídos por estarem fora do escopo, repetidos entre uma base de dados e outra, e alguns com acesso restrito. Nota-se que a quantidade de trabalhos selecionados da IEEE *Xplore* foi maior, visto que foi a primeira a ser analisada e não foram encontrados trabalhos repetidos; já nas demais bases este fator influenciou no número de artigos encontrados e inclusos na revisão. Deste modo, dentro da base IEEE *Xplore* foi aproveitado cerca de 24,14% dos artigos, na *Scopus* 19,05%, *ScienceDirect* 9,1% e na *PubMed* 18,2% dos trabalhos foram utilizados. Assim, prosseguiu-se com a avaliação dos artigos considerando-se as questões norteadoras desta revisão.

**Séries temporais -** A série temporal (*time series*) é criada a partir de observações em ordem cronológica [7]. De forma geral, uma série temporal *s* pode ser descrita por meio de uma sequência ($x_1$, $x_2$, ..., $x_n$) tendo *n* pontos de informações $x_i$. Esses pontos de informações podem ser números reais, valores de tensão ou corrente, temperatura ou dados multidimensionais [16]. Além disso, uma série temporal pode conter um grande volume de dados, alta dimensionalidade e necessitar de continuas atualizações das informações. No geral, uma série temporal reduz a dimensão dos dados originais [7], e ela pode conter valores numéricos contínuos e precisa ser transformada e/ou discretizada ajustando os dados para a mineração [16].

No entanto, a discretização pode distorcer a série temporal amostrada ou comprimida, caso a taxa de amostragem seja muito baixa [7]. O método *Symbolic Aggregate Approximation* (SAX) e os baseados em agrupamentos são os mais utilizados para discretizar uma série temporal com pontos equidistantes no tempo [16]. Para este mesmo propósito tem-se outras técnicas como: *Segmented Sum Variation (SSV) e multiple abstraction level mining (MALM)* proposto por Li et al. (1998) apud [7], *Perceptually Important Points (PIP)* por Chung et al. (2001) apud [7], todas no domínio do tempo. Para representar uma série temporal no domínio da frequência, pode-se usar: *discrete Fourier transforms (DFT), discrete wavelet transform (DWT) e Haar transform mostrados por* Chan e Fu (1999) apud [7], *Principal Component Analysis (PCA) e* Modelos Ocultos de Markov (*HMM)* citados por Azzouzi and Nabney (1998) apud [7]. Após a discretização, pode-se aplicar a mineração na série temporal, onde objetiva-se buscar por padrões de similaridade, tendências, mutações ou *outliers* auxiliando na tomada de decisão [7].

**Mineração de dados -** A mineração de dados (*data mining*) visa explorar e analisar grandes volumes de dados de forma automática ou semiautomática. Minerar dados é o objetivo final para descobrir informações a partir de dados originais ou alterados de uma série temporal [7], e também, para encontrar informações desconhecidas dentro de uma base de dados confiável [22]. O reconhecimento de padrão é uma tarefa comum a ser feita e, após esta etapa, é realizada a classificação, depois associação de regras de mineração e, por fim, a visualização [7] [23].

A busca de padrões varia de acordo com o interesse e o domínio que se encontram os dados, já o agrupamento é um processo de aglutinar padrões encontrados baseados num dado critério de similaridade [8]. Os métodos de agrupamento podem ser divididos em: particionamento, métodos hierárquicos, baseados em: densidade, estruturas em *grid* e/ou modelos [25]. Na classificação é comum realizar pré-processamento de dados, ou seja, pode-se combinar propriedades ou usar padrões identificados na série temporal. Já a associação de regras é focada em itens simbólicos presentes nas transações. A árvore de decisão também é utilizada para definir regras de mineração. Por fim, a visualização visa delinear uma característica para um subconjunto de dados [4] [7].

O algoritmo *Motif Discovery* visa encontrar padrões desconhecidos em uma série temporal sem qualquer dado prévio sobre seus locais ou formas [17]. A análise de regressão e k-*nearest neighbor* (KNN) são métodos muito usados, o KNN possui alta precisão e não apresenta rigor na parametrização dos dados. A árvore de decisão classifica os dados de classes desconhecidas, já a rede Bayesiana é baseada em conhecimento probabilístico e usada em previsões [22]. A compreensão destes conceitos é de suma importância visto o enorme volume de dados a serem lidos e/ou escritos. Assim, a mineração de dados em séries temporais possui inúmeras aplicações diretas dentro da indústria, análise de dados bancários, análise de fraudes, mineração de bio-dados, análise de dados de assistência médica etc., auxiliando na tomada de decisão [4] [6].

**Tabela 2** – Artigos e métodos utilizados para mineração de dados relacionados à doença de Parkinson (DP).

| Artigo | Ano | Objetivo | Métodos de mineração empregados | Resultado | Possui relação com a DP? |
|---|---|---|---|---|---|
| [1] | 2013 | Revisar os algoritmos de monitoramento fisiológico, a partir de sensores móveis, de sinais vitais: eletrocardiogramas, saturação de oxigênio, batimentos cardíacos, glicose no sangue, taxa de respiração e pressão sanguínea. | *Support Vector Machines*, Rede Neural, transformada *Wavelet*, ferramentas estatísticas, *Hidden Markov Models* e árvore de decisão | O artigo revela os algoritmos mais usados para cada tipo de sinal fisiológico monitorado. | Não |
| [3] | 2007 | Classificação de dados em série temporal para diagnosticar a atividade cerebral relacionada à epilepsia usando o eletroencefalograma (EEG). | *K-Nearest Neighbor* | Sensibilidade de 81,29% e especificidade de 72,86% na classificação. | Não |
| [2] | 2008 | Comparar o agrupamento das atividades cerebrais normais com as atividades pré-convulsão epilética. | *Clustering via concave quadratic programming* (CCQP) | A comparação das atividades cerebrais foi alcançada. | Não |
| [12] | 2010 | Utilizar o método ST-ICA para extrair informações espaço-tempo derivadas de multicanais de dados provenientes de EEG. | *Independent component analysis* (ICA). | A demonstração do uso do ST-ICA foi explicada em dados de EEG. | Não |
| [13] | 2010 | Usar técnicas de mineração de dados para avaliar a pressão intracraniana de forma não invasiva (NICP). | *Ordinary Least Squares*, *Recursive Weighted Least Squares*, *Kernel Spectral Regression*, Rede neural e *local linear fuzzy* | O algoritmo KSR obteve maior precisão, porém poucos dados foram utilizados. | Não |
| [10] | 2016 | Pré-processar uma série temporal e realizar o reconhecimento do padrão para gerar sequências temporais da série original e depois extrair padrões temporais da sequência. | *Piecewise Aggregation Aproximation* (PAA) e identificação de picos. | Foi mostrada a extração temporal de sequências, mas não foi testado em dados reais. | Não |
| [24] | 2016 | Analisar os dados de reivindicação de seguros de saúde para prever tempo de internação. | Árvore de decisão e coeficiente de correlação Matthews (MCC). | O modelo semestral obteve melhor desempenho. | Não |
| [15] | 2010 | Reconhecer anomalias por meio da investigação da marcha em indivíduo com hemiplegia, Parkinson, com dores nas costas e dores nas pernas. | *naive Bayes e* redes neurais artificiais. | 97,2% com o classificador *naive Bayes* e 100% ao utilizar redes neurais | Sim |
| [14] | 2010 | Utilizar pseudo-séries temporais a partir de dados de corte transversal para identificar regiões de interesse relacionadas às doenças: glaucoma, Parkinson e câncer de mama. | Redes Bayesianas dinâmicas e *Hidden Markov Models (*HMM). | Uma previsão média de 94,2% foi alcançada, viabilizando o corte transversal. | Sim |
| [9] | 2011 | Analisar o tremor a partir de métodos de particionamento usados nos resultados de análise do tratamento via estimulação profunda do cérebro (DBS) | *k-means* e *k-medoids* | Os parâmetros não lineares não classificaram o tremor | Sim |
| [20] | 2011 | Analisar os movimentos associados à medicação Levodopa aplicada em pacientes com a doença de Parkinson utilizando a arquitetura epiNet | Arquitetura epiNet baseada em processos genéticos e epigenéticos. | No treinamento, obteve-se 80% de precisão, já nos testes 70%. | Sim |
| [5] | 2014 | Usar *Joint blind source separation* (JBSS) no banco de dados contendo EEG multidimensional e dados cinemáticos unidimensional utilizando EEMD-IVA. | *Ensemble Empirical Mode Decomposition*, *Independent Vector Analysis* | Discriminação realizada entre indivíduos com Parkinson e de controle. | Sim |
| [19] | 2017 | Estimar o índice Hoehn & Yahr a partir dos dados dos sintomas motores utilizando a escala UPDRS em pacientes com a doença de Parkinson | Árvore de decisão e *Repeated Incremental Pruning to Produce Error Reduction* | As abordagens atingiram uma precisão em torno de 87% de acerto. | Sim |

## Discussão

Após conceituar as séries temporais percebe-se que este formato de disposição da informação em ordem cronológica é muito utilizado. Já a mineração de dados é indispensável para automatizar e ajudar na busca por conhecimento útil. Neste sentido, a correlação entre série temporal e mineração de dados é quase simbiótica.

A Tabela 2 apresenta os trabalhos analisados, relacionando as técnicas mais usadas da mineração de dados dentro da área da engenharia biomédica. Assim, as redes neurais citadas em [1], [13] e [15], junto com a árvore de decisão em [1], [19] e [24] foram as mais utilizadas. As redes Bayseianas [14] e [15] e os Modelos Ocultos de Markov [1] e [14] também se destacaram.

Ainda na Tabela 2, pode-se perceber que de todos os trabalhos apenas 6 tinham relação com a doença de Parkinson [15], [14], [9], [20], [5] e [19]. E destes 6 trabalhos, 4 relatavam somente a doença de Parkinson [9], [20], [5] e [19], e, por fim, dos 4 últimos trabalhos apenas 1 utilizou dados do tremor de pacientes [9]. Deste modo, fica evidente que a mineração de séries temporais em dados de tremor de pacientes com Parkinson carece de investigações mais detalhadas.

## Conclusão

Apesar do trabalho elucidar os conceitos de séries temporais e mineração de dados mostrando as técnicas mais usuais dentro da Engenharia Biomédica, e também, evidenciar a existência de poucos trabalhos relacionando a mineração de séries temporais com a doença de Parkinson, futuramente, pode-se aprimorar o processo de busca, por meio da construção de novas *strings* e investigação em bases de dados adicionais.

## Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). A. O. Andrade é Bolsista do CNPq, Brasil (305223/2014-3).

**Referências**


[1] BANAEE, H.; AHMED, M. U.; LOUTFI, A. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Switzerland, v.13, n.12, p.17472–17500, 2013.

[2] CHAOVALITWONGSE, W. A. Novel quadratic programming approach for time series clustering with biomedical application. Journal of Combinato- rial Optimization, v.15, n.3, p.225–241. 2008.

[3] CHAOVALITWONGSE, W. A.; FAN, Y. J.; SACHDEO, R. C. On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, v.37, n. 6, p.1005–1016, 2007.

[4] CHEN, F. et al. Data Mining for the Internet of Things: Literature Review and Challenges. International Journal of Distributed Sensor Networks, v.11, n.8, p.1, 2015.

[5] CHEN, X. et al. An EEMD-IVA framework for concurrent multidimensional EEG and unidimensio- nal kinematic data analysis. IEEE Transactions on Biomedical Engineering, v.61, n.7, p.2187–2198, 2014.

[6] DAS, A. K. et al. Data mining techniques in Indian healthcare: A short review. International Conference on Man and Machine Interfacing. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7456611/> Acesso em: 12 jun. 2017

[7] FU, T. C. A review on time series data mining Engineering Applications of Artificial Intelligence, 2011. Disponível em: <http://linking hub.elsevier.com/retrieve/pii/S095219761000172. Acesso em: 6 jun. 2017

[8] GARIMA; RANI, S. Review on time series databases and recent research trends in Time Series Mining. International Conference - The Next Generation Information Technology Summit. 2014. Disponível em: <http://ieeexplore.ieee.org/document/6949290/ >. Acesso em: 12 jun. 2017

[9] GEMAN, O.; TURCU, C. Partitioning methods used in DBS treatments analysis results. International Conference on Neural Networks. 2011. Disponível em: <http://ieeexplore.ieee.org/document/6033441 />. Acesso em: 13 jun. 2017.

[10] INIBHUNU, C.; MCGREGOR, C. Machine Learning Model for Temporal Pattern Recognition. 2016. IEEE EMBS International Student Conference, p. 1–4, 2016.

[11] INTERNET LIVE STATS. Real Time Statistics Project. 2016. Disponível em: <http://www.intern etlivestats.com/>. Acesso em: 20 de Jan. 2017.

[12] JAMES, C. J.; DEMANUELE, C. Space-Time Independent Component Analysis: The definitive BSS technique to use in biomedical signal processing? Annual International Conference of the IEEE Engineering in Medicine and Biology. 2010. Disponível em: <http://ieeexplore.ieee.org/ document/5627351/>. Acesso em: 13 jun. 2017

[13] KIM, S. et al. Noninvasive Intracranial Pressure Assessment based on Data Mining Approach using Nonlinear Mapping Function. IEEE Transactions Biomedical Engineering, v. 59, n. c, p. 1–7, 2010.

[14] LI, Y.; TUCKER, A. Uncovering disease regions using pseudo time-series trajectories on clinical trial data. 3rd International Conference on Biomedical Engineering and Informatics, 2010. Disponível em: <http://ieeexplore.ieee.org/ document/5639726/>. Acesso em: 13 jun. 2017.

[15] POGORELC, B.; GAMS, M. Discovery of gait anomalies from motion sensor data. International Conference on Tools with Artificial Intelligence. 2010. Disponível em: <http://ieeexplore.ieee.org/ document/5670088/>. Acesso em: 13 jun. 2017

[16] SCHLUTER, T.; CONRAD, S. About the analysis of time series with temporal association rule mining. Computational Intelligence and Data Mining, IEEE Symposium, p. 325–332, 2011.

[17] TORKAMANI, S.; LOHWEG, V. Survey on time series motif discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v.7, n.2, p.1–8. 2017.

[18] SUN, Jimeng; REDDY, Chanddan K. Big Data Analytics for Healthcare: Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013. Disponível em: <http:// dmkd.cs.vt.edu/TUTORIAL/Healthcare/>. Acesso em: 13 de janeiro de 2017.

[19] TSIOURIS, K. M. et al. Mining Motor Symptoms UPDRS Data of Parkinson's Disease Patients for the Development of Hoehn and Yahr Estimation Decision Support System. IEEE International Conference on Biomedical & Health Informatics, p. 445–448, 2017.

[20] TURNER, A. P. et al. Using epigenetic networks for the analysis of movement associated with levodopa therapy for Parkinson's disease. BioSystems, v. 146, p. 35–42, ago. 2016.

[21] ULARU, E. G. et al. Perspectives on Big Data and Big Data Analytics. Database Systems Journal, v. III, n. 4, p. 3–14, 2012.

[22] WANG, S. A comprehensive survey of data mining-based accounting-fraud detection research. International Conference on Intelligent Computation Technology and Automation, 2010. Disponível em: <http://ieeexplore.ieee.org/ document/5522816/>. Acesso em: 12 jun. 2017

[23] WILSON, S. J. Data representation for time series data mining: time domain approaches. Wiley Interdisciplinary Reviews: Computational Statistics, v. 9, n. 1, p.1–6, 2017.

[24] XIE, Y. et al. Analyzing health insurance claims on different timescales to predict days in hospital. Journal of Biomedical Informatics, v. 60, p. 187–196, 2016.

[25] ZOLHAVARIEH, S.; AGHABOZORGI, S.; TEH, Y. W. A review of subsequence time series clustering. Scientific World Journal, v. 2014, p. 1-19, 2014.