

Algumas Abordagens para Mineração de Séries Temporais aplicadas a Dados de Segurança de Barragens *

André Gustavo Maletzke^{1,2}, Huei Diana Lee¹,
Maria Carolina Monard², Wu Feng Chung^{1,3}

¹Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

²Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

³Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

{andregustavom, hueidianalee}@gmail.com

Abstract. *Several real-world applications deal with large amounts of temporal data, which must be treated accordingly by the data mining process in order to model and extract information from this sort of data. Conventionally, temporal data is classified to either categorical event streams or numerical time series, and both types have been studied in data mining and statistics. This work presents some techniques developed to deal with the irregularity of temporal data such that the data mining process takes into account the temporal information of the data. These techniques will be applied to real data obtained from monitoring procedures related to dams' security area.*

Resumo. *Inúmeras aplicações manipulam grandes quantidades de dados temporais, os quais podem ser tratados por processos de mineração de dados para modelar e extrair informações desses dados. Geralmente, dados temporais são classificados como eventos categóricos ou numéricos, sendo que ambos têm sido estudados tanto em mineração de dados quanto em estatística. Este trabalho apresenta algumas abordagens aplicadas na área mineração de dados aplicada à análise de séries temporais. Essas abordagens serão aplicadas em dados reais obtidos a partir de processos de monitoração relacionados à área segurança de barragem.*

1. Introdução

O advento de novas tecnologias tem impulsionado o acúmulo de informações providas de distintas fontes como medicina, economia e segurança. Esse fato tem contribuído para o surgimento de grandes Bases de Dados — BD —, tornando o processo de análise e

*Trabalho desenvolvido com o apoio da Fundação Parque Tecnológico Itaipu – FPTI

interpretação desses dados uma tarefa crescentemente complexa aplicando-se métodos e ferramentas tradicionais. Assim, torna-se necessário o desenvolvimento e a aplicação de novos processos e técnicas computacionais para realizar essa tarefa. Um dos processos de apoio à análise de dados é o de Descoberta de Conhecimento em Bases de Dados — DCBD — [Fayyad et al. 1996]. Esse processo possibilita a extração de padrões contidos nesses dados com o intuito de que possam auxiliar especialistas de diversos domínios no processo de tomada de decisão.

Em muitos domínios, uma parte dos dados provem de processos de monitoração realizados ao longo do tempo, seja de modo contínuo ou equi-espaçado, caracterizando um tipo de dado denominado de Série Temporal — ST. Uma série temporal pode ser definida como uma coleção de observações de um determinado fenômeno realizadas ao longo do tempo [Morettin and Toloí 2006]. Portanto, ao contrário do que ocorre na maioria dos problemas de DCBD, em séries temporais a ordem dos dados é crucial para a análise [Last et al. 2004]. Desse modo, esses dados devem ser pré-processados, por meio de técnicas de Análise de Series Temporais — AST — para que possam ser utilizados pelo processo de DCBD.

Um dos temas que tem despertado interesse e preocupações de inúmeras instituições e profissionais nacionais e internacionais refere-se à Segurança de Barragens — SB. O crescente interesse por esse tema é motivado pela necessidade de se tomar decisões para corrigir deficiências que possam, eventualmente, ameaçar a segurança de barragens, bem como ampliar a conscientização, a organização e a colaboração das entidades envolvidas. Portanto, a avaliação de riscos de uma barragem deve possibilitar a identificação de problemas e recomendar reparos corretivos, restrições operacionais e/ou modificações quanto às análises e aos estudos para determinar as soluções para eventuais problemas [Menescal et al. 2001]. As avaliações realizadas são baseadas, normalmente, em um conjunto de dados que são adquiridos por meio de processos de monitoração e análises mecânicas de diversos materiais, os quais podem gerar um conjunto considerável de observações ordenadas no tempo, isto é, séries temporais.

Nesse contexto, o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP — e o Laboratório de Bioinformática — LABI — da Universidade Estadual do Oeste do Paraná — UNIOESTE — em parceria com a Itaipu Binacional por meio do Centro de Estudos Avançados em Segurança de Barragens — CE-ASB, estão desenvolvendo uma metodologia de identificação, avaliação e monitoramento de eventos relacionados ao tema de segurança em barragens. Neste trabalho é apresentado um estudo preliminar no qual são abordados alguns dos métodos presentes na literatura para Mineração de Dados — MD — aplicados à mineração de séries temporais. Esse estudo faz parte de um projeto mais amplo cujo objetivo é a extração de conhecimento existente nos dados relacionados ao processo de monitoração de barragem em usinas hidrelétricas.

Este trabalho está estruturado da seguinte maneira: na Seção 2 são apresentadas algumas abordagens para mineração de dados aplicada à ST; na Seção 3 é apresentada a discussão do trabalho e na Seção 4 são apresentadas as conclusões deste trabalho e trabalhos futuros.

2. Abordagens para Mineração de Séries Temporais

Métodos tradicionais de mineração de dados são, normalmente, aplicados sobre bases de dados estáticas, ou seja, BD nas quais a ordem dos exemplos não tem influência para o processo de reconhecimento de padrões. Entretanto, existem aplicações nas quais essa característica sequencial dos dados é importante no processo de mineração [Last et al. 2001], o que tem aumentado o interesse pela mineração de séries temporais. Uma quantidade considerável de trabalhos científicos foram desenvolvidos mostrando a aplicação de métodos tradicionais para mineração de ST, bem como propondo novos métodos para esse fim [Khatkhate et al. 2006, Wei and Keogh 2006, Keogh and Kasetty 2002, Last et al. 2001]. Grande parte dos métodos para a análise inteligente de ST destinam-se basicamente às tarefas de predição, segmentação, agrupamento, indexação, classificação, identificação de novidade e visualização [Keogh and Lin 2005]. Neste trabalho são descritas as abordagens referentes a três dessas tarefas, pois estão diretamente relacionadas com a sua aplicação a dados provenientes de processos de monitoração de barragens. São elas:

- Predição: dada uma série temporal com um determinado comportamento tem por objetivo realizar predições futuras do comportamento dessa ST;
- Agrupamento (*Clustering*): a partir de uma base de dados de ST tem por objetivo encontrar agrupamentos naturais entre os elementos da base sobre uma determinada medida de similaridade;
- Classificação: dada uma ST não rotulada o objetivo é designar uma classe a essa série por meio de classes pré-definidas.

Predição

Diversas abordagens foram propostas e avaliadas para realizar predição de ST [Chong and Yoo 2006, Nguyen and Chan 2004, Sorjamaa et al. 2005, Liu and Liu 2002]. Geralmente, a tarefa de se realizar predições futuras consiste em construir modelos que a partir de acontecimentos já conhecidos possam realizar predições de acontecimentos futuros. Na Figura 1 é apresentada uma representação esquemática ilustrando uma série temporal (linha contínua) e a predição de valores futuros dessa série (linha tracejada). Existem diversos métodos para a construção desses modelos, dentre os quais os modelos estatísticos são amplamente utilizados [Morettin and Toloí 2006]. Dentre esses modelos destacam-se os modelos de Regressão Polinomial — RP —, Auto-regressivos — AR — e Auto-regressivos Integrados de Médias Móveis — ARIMA — [Latorre and Cardoso 2001]. No entanto, esses modelos possuem algumas limitações como linearidade e outros requisitos que envolvem características de uma ST.

Outros modelos são baseados em métodos de Aprendizado de Máquina — AM. Na área de AM existem inúmeros métodos

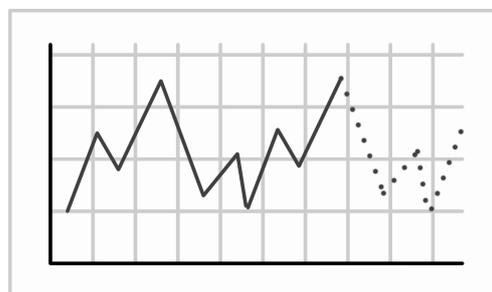


Figura 1. Representação esquemática da predição de valores futuros de uma ST.

destinados, por exemplo, a identificação de padrões, a aquisição de novos conhecimentos e a predição de eventos futuros utilizando um conjunto de dados (exemplos) de treinamento que descrevem eventos do passado. Porém, esses modelos são, normalmente, aplicados a bases de dados estáticas na qual a ordem dos acontecimentos não apresenta nenhuma importância para a tarefa de predição. Para isso, esses modelos devem ser adaptados ou características de uma ST devem ser extraídas para que tais modelos possam ser aplicados.

Em [Last et al. 2001] é proposta uma metodologia aplicada ao processo de DCBD de séries temporais utilizando AM. A metodologia está dividida em três etapas: pré-processamento de ST, mineração de dados e pós-processamento. Na etapa de pré-processamento são realizadas as tarefas de limpeza dos dados e extração de características. Inicialmente é realizada a limpeza dos dados por meio da identificação e redução de ruídos presentes em uma ST. Posteriormente, é realizada a extração de características de uma ST. As características consideradas nesse trabalho são baseadas em tendência e ruído.

Na etapa de mineração de dados é realizada a construção do modelo, extração de regras de associação e redução de regras. Um dos métodos que pode ser aplicado para construção do modelo é baseado no método denominado *Information-Theoretic* [Maimon et al. 1999]. Assim, as iterações entre os atributos de entrada e os atributos classe são modeladas por uma *Information-Theoretic Connectionist Network*, a qual consiste de um nó raiz, um número de camadas intermediárias e uma camada de saída. Os nós da cada camada intermediária são associados com os diferentes valores dos atributos de entrada. Esse modelo é baseado em um método de AM amplamente conhecido, denominado de Rede Neural Artificial — RNA —, a qual consiste de construções matemáticas simplificadas inspiradas no sistema nervoso humano [Haykin 1999]. O modelo proposto em [Last et al. 2001] é utilizado na construção de regras de associação que descrevem uma dada ST, o qual pode ser usado para prever valores desconhecidos do atributo classe, de modo similar à classificação utilizando uma árvore de decisão.

No entanto, um dos grandes desafios na predição de ST é a realização da predição de uma grande quantidade de eventos futuros [Nguyen and Chan 2004]. Em [Sorjamaa et al. 2005] é proposta uma metodologia para a predição de um período longo de uma série temporal baseada em duas estratégias de predições denominadas *Recursive Prediction* — RP — e *Direct Prediction* — DP. A estratégia RP utiliza valores já preditos para prever os próximos valores. Já a estratégia DP realiza predições utilizando-se de valores conhecidos *a priori*. Essa estratégia é considerada mais complexa se comparada com a primeira, no entanto, o resultado das predições são considerados mais precisos. Para a construção do modelo são consideradas três abordagens para a seleção do conjunto de entrada ao modelo: *k-Nearest Neighbor*, *Mutual Information* — MI — e *Nonparametric Noise Estimation* — NNE [Mitchell 1997]. Desse modo, o conjunto de entrada para o modelo é o que obtém o melhor dos três critérios. Para a construção do modelo de predição é utilizado um método baseado em *Support Vector Machine*, o qual é inspirado na teoria de aprendizado estatístico. A ideia principal de um modelo SVM é obter o hiperplano que separa melhor os exemplos em função do erro do modelo [Pal and Mitra 2004]. Mais especificamente, o método utilizado em [Sorjamaa et al. 2005] é o *Least Squares Support Vector Machine* — LS-SVM — o qual possibilita construir um modelo que evita de modo eficiente o problema de mínimos

locais [Lendasse et al. 2004].

Encontra-se na literatura uma ampla quantidade de trabalhos que utilizam os métodos baseados no paradigma conexionista de AM para realizar previsões de ST [Chong and Yoo 2006, Nguyen and Chan 2004, Crone 2002]. No entanto, a utilização de um modelo conexionista para realizar a previsão de eventos futuros longos de uma série tem se mostrado bastante ineficiente quando utilizados métodos de construção desses modelos. Nesse sentido, uma nova abordagem para a construção tradicional de modelos baseados em redes neurais artificiais é proposta por [Nguyen and Chan 2004]. Nessa abordagem é utilizada mais de uma RNA, as quais trabalham em conjunto para realizar a tarefa de previsão de acontecimentos futuros. Esse modelo mostrou-se mais eficiente se comparado a modelos baseados em uma única RNA.

Agrupamento (Clustering)

Uma série temporal pode ser considerada como um conjunto amplo de pontos em um plano, de modo que tal conjunto possa ser considerado como um único objeto. Desse modo, uma das áreas que tem motivado um número considerável de pesquisas está relacionada à capacidade de, dado um conjunto de eventos, poder realizar uma série de agrupamentos desses objetos de modo a descrever esse conjunto por meio desses agrupamentos. Na área de análise de ST, técnicas de agrupamento são amplamente estudadas e aplicadas a domínios como medicina e indústria [Latorre and Cardoso 2001, Liu and Liu 2002].

Em [Keogh and Lin 2005] a tarefa *clustering* aplicada a dados de séries temporais é classificada em duas categorias:

- *Whole clustering*: nessa categoria a idéia de *clustering* é semelhante a aplicada sobre dados discretos, sendo o principal objetivo agrupar ST similares dentro de um mesmo *cluster*.
- *Subsequence clustering*: nesse caso, a partir de uma ST são extraídas subsequências dessa série por meio de uma técnica denominada de janela deslizante (*sliding window*). Desse modo, o *cluster* é construído unicamente a partir das subsequências extraídas e não mais da ST original.

Em [Wang et al. 2006] é proposta uma abordagem para se realizar *clustering* de séries temporais baseada em características estruturais. Uma das justificativas para utilizar uma abordagem baseada em características é devido a alta dimensionalidade de uma ST, o que faz com que o processo de MD seja menos eficiente. A proposta é baseada nos métodos mais tradicionais de agrupamento de ST, os quais utilizam-se de medidas de distância. Diversos métodos de extração de características são propostos na literatura voltados para os mais variados propósitos. Uma ST pode ser descrita por um conjunto de termos qualitativos como sazonalidade, tendência e ruído. Entre as características podem ser utilizadas: *Trend*, *Seasonality*, *Serial Correlation*, *Non-linearity*, *Skewness*, *Kurtosis*, *Self-similarity*, *Chaotic*, *Periodicity*.

O método de agrupamento utilizado em [Wang et al. 2006] é o *Self-Organizing Maps* — SOM [Pal and Mitra 2004]. Esse método trabalha de modo similar ao *K-means*, o qual permite que seja escolhido um número *K* de *clusters* a serem criados a partir dos dados. Porém, o SOM permite a escolha da forma e do tamanho da rede de *clusters* a ser construída a partir desses dados. Portanto, no SOM cada *cluster* é denominado de nó, sendo que normalmente esses nós são dispostos em um formato similar ao de uma

rede, na qual a largura e a altura devem ser definidas como parâmetros. Na Figura 2 (a) é apresentado um exemplo de dendograma construído aplicando o método SOM.

Em [Keogh et al. 2004] é proposto um método para comparar ST de alta dimensionalidade usando *Co-compressibility* como uma medida de dissimilaridade, denominado *Compression-based Dissimilarity Measure* — CDM. Essa medida é baseada no conceito de *Kolmogorov complexity* e pode ser utilizada por algoritmos de MD como os algoritmos de *clustering*. Existem outras medidas de similaridade amplamente utilizadas em diversos domínios como distância Euclidiana, *Uniform Scaling* — US — e *Dynamic Time Warping* — DTW [Morettin and Toloí 2006]. A medida DTW permite realizar a aproximação de séries que possuem desenvolvimentos semelhantes, porém defasados ou deformados um em relação ao outro, sobre o eixo do tempo [Morettin and Toloí 2006, Last et al. 2004]. O ponto forte dessa medida deve-se ao fato de que as comparações entre os pares não são estritamente realizadas nas mesmas coordenadas, como acontece por exemplo com a distância Euclidiana, mas com uma flexibilidade de que um ponto possa ser comparado com pontos adjacentes ao de outra série.

Ainda em [Keogh et al. 2004] é apresentado um estudo comparativo entre a tarefa de *clustering* utilizando CDM e Distância Euclidiana. Na Figura 2 (b) é ilustrado o dendograma gerado para as mesmas bases de dados porém variando a medida de similaridade utilizada.

Em [Fu et al. 2005] é proposta uma nova medida para comparação de séries temporais baseada na medida DTW e na US, denominada *Scaling and Time Warping* — SWM. A SWM é uma medida que combina os efeitos da DTW e da US, sendo uma alternativa híbrida a ser aplicada em problemas nos quais tanto a US quanto a DTW, aplicadas isoladamente, não obtém resultados significativos.

Classificação

Problemas que envolvem classificação têm atraído grande interesse da comunidade científica nas últimas décadas. Nesse contexto, o problema de classificação de séries temporais cada vez mais vem sendo discutido e estudado de modo que possa ser utilizado em aplicações reais, tais como na classificação de documentos manuscritos, bases de dados de vídeos e em bases de dados de exames de Eletrocardiograma. Em [Wei and Keogh 2006] é apresentada uma proposta de utilização de aprendizado semi-supervisionado aplicado a base de dados de ST.

Normalmente, o método mais utilizado para realizar a classificação de uma ST é o

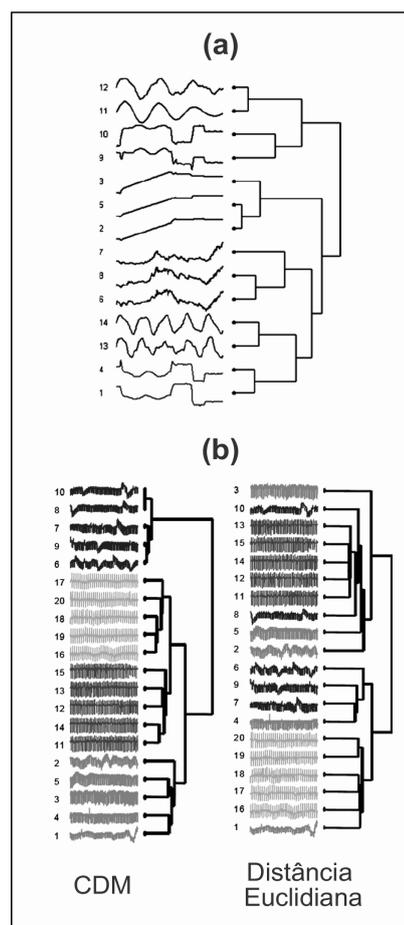


Figura 2. Representação esquemática da geração de Clusters.

algoritmo *One-Nearest-Neighbor* — ONN — [Wei and Keogh 2006, Keogh et al. 2004, Last et al. 2004, Keogh and Kasetty 2002]. No entanto, vários outros algoritmos foram propostos para o problema de classificação. Entretanto, o ONN utilizando distância Euclidiana mostrou-se melhor entre todos os demais [Keogh and Kasetty 2002].

Em [Geurts 2001] assume-se como hipóteses que o problema de classificação de uma ST deve ser resolvido por meio da identificação e da combinação de características de uma ST. Posteriormente, esses padrões são combinados e um classificador é construído, de modo que regras de classificação possam ser extraídas. Na construção desse modelo é utilizada uma abordagem híbrida entre os métodos de classificação binária e de árvore de decisão.

Uma abordagem de representação de exemplos amplamente utilizada na área de AM para a indução de modelos é a representação atributo-valor. Nesse abordagem a base de dados é composta por um conjunto de tuplas, na qual cada tupla representa um exemplo descrito por um conjunto de atributos. Nesse sentido, com intuito de aplicar essa abordagem para a indução de modelos, [Kadous and Sammut 2004] propõe a utilização de *metafeatures*, as quais são expandidas a um determinado domínio de modo que seja possível a construção de atributos a partir de ST. Desse modo, nesse trabalho um conjunto genérico de *metafeatures* foi capaz de gerar um classificador compreensível e eficiente. Os experimentos realizados nesse trabalho utilizaram bases de dados de Eletrocardiograma e mostraram-se satisfatórios.

3. Discussão

A Usina Hidrelétrica de Itaipu¹ é um empreendimento binacional desenvolvido por meio de uma parceria entre Brasil e Paraguai, que tem por objetivo suprir as necessidades energéticas e fomentar o desenvolvimento de ambas as nações. Atualmente é a maior usina em produção energética do mundo, sendo considerada como referência em produção de energia. Além disso, é considerada como um marco da engenharia, conformada por aproximadamente 13,5 milhões de m^3 de concreto. Por meio das suas características, constitui-se como fonte de geração de conhecimento e pesquisa nos temas referentes à construção, manutenção e segurança de barragens.

Vários controles realizados na usina são baseados, normalmente, em um conjunto de dados que são adquiridos por meio de processos de monitoração e análises mecânicas de diversos materiais, os quais podem gerar um conjunto considerável de observações ordenadas no tempo. Atualmente, Itaipu possui centenas de sensores que capturam inúmeras informações como temperatura, umidade, vazão, deslocamento e pressão. Outros dados são coletados por meio de análises provenientes, por exemplo, de testes físico-químicos e biológicos de materiais da barragem. Desse modo, esses dados representados por ST podem ser estudados por meio de métodos de AST com o objetivo de se extrair conhecimento desses dados, de tal maneira que esse conhecimento possa ser utilizado no processo de identificação, avaliação, prevenção e tratamento de riscos.

Uma das tarefas que pretende ser realizada a partir desses dados é a de predição, na qual espera-se que dados que descrevam o histórico comportamental, por exemplo, do nível da barragem nos últimos anos possa ser utilizado para identificar com maior

¹<http://www.itaipu.gov.br>

precisão a variação de nível, por exemplo, nas próximas semanas, podendo auxiliar no controle de nível da barragem de modo mais eficiente. Nesse sentido, a metodologia proposta em [Sorjamaa et al. 2005], apresentada anteriormente, mostra-se interessante e, possivelmente, aplicável a esse caso, pois devido ao fato de serem dados de monitoração constante a quantidade de amostras coletadas é consideravelmente grande, tornado-se necessário realizar-se a predição de um longo período da ST que representa esses dados.

Outra abordagem relevante é a utilização de um modelo baseado em uma RNA, no entanto, como mencionado, a construção desse modelo para essa finalidade tem se mostrando pouco eficiente. Porém, em [Nguyen and Chan 2004] é adotada uma abordagem baseada na utilização de várias RNAs atuando em conjunto para se realizar a tarefa de predição. Essa abordagem poderá também ser aplicada ao problema apresentado neste trabalho.

Consideramos que a tarefa de classificação de ST será de grande utilidade no contexto de segurança em barragens, por exemplo, com intuito de identificar um comportamento anômalo em máquinas da usina. Os dados coletados em relação ao desempenho de uma máquina ao longo do tempo poderão demonstrar se ela está operando em condições normais ou apresenta algum tipo de fadiga. Ressalta-se que a tarefa de classificação tem despertado o interesse de diversas áreas, por exemplo, em [Gupta et al. 2005] são aplicados conceitos de AST para detectar danos causados pela fadiga de equipamentos. De modo similar, métodos de classificação também poderão ser aplicados ao tema de SB.

Como mencionado, outra tarefa comumente realizada, principalmente, quando a BD é conformada por exemplos não rotulados, é a de *clustering*. Na área de AST essa tarefa é amplamente aplicada, pois é frequente a não existência de classe associada à ST. Esse fato é um dos problemas encontrados na área de classificação, na qual há uma escassez de dados rotulados que possam ser usadas na indução de classificadores. Desse modo, a abordagem de *clustering* pode auxiliar tanto na construção de BD rotuladas quanto na descrição de um conjunto de dados de modo que o especialista possa identificar, visualmente, com maior facilidade padrões contidos nesse dados.

4. Conclusões

Nas últimas décadas têm ocorrido uma explosão na quantidade de dados que são registrados diariamente, fato motivado pelo desenvolvimento de tecnologias digitais que suportam um acúmulo considerável de dados. No entanto, somente na última década surgiu um interesse maior por áreas de pesquisas que envolvem o estudo dos dados e sua relação com o tempo, buscando-se modelar essa relação. Atualmente, a área de AST encontra-se em crescente acensão, pois cada vez mais busca-se estudar o comportamento dos dados ao longo do tempo. Neste trabalho foram apresentadas algumas abordagens que buscam resolver problemas nas áreas de predição, agrupamento e classificação.

Como mencionado, este trabalho está inserido no Projeto de Análise Inteligente de Dados no Centro de Estudos Avançados em Segurança de Barragens. Nesse projeto estarão colaborando as áreas de Engenharia, Biologia, Física e Inteligência Computacional para o desenvolvimento de uma metodologia de identificação, avaliação e monitoramento de eventos relacionados ao tema de SB. Neste trabalho foi apresentado um estudo inicial sobre algumas abordagens existentes na literatura relacionadas ao processo de mineração de séries temporais. Esse estudo servirá de base para trabalhos futuros, os quais têm

como objetivo a identificação de padrões que possam estar contidos em dados oriundos de processo de monitoração, análise periódica de materiais, entre outros, obtidos da Usina Hidrelétrica de Itaipu. Espera-se que futuramente a aplicação de métodos de IA sobre os dados da barragem possa auxiliar na avaliação e estimação de possíveis riscos, recomendar reparos corretivos, restrições operacionais e/ou modificações quanto às análises e estudos para determinar as soluções para eventuais problemas com o intuito de auxiliar a especialistas e pesquisadores no processo de tomada de decisão.

Este trabalho, por tratar-se de um estudo inicial sobre algumas abordagens aplicadas à mineração de séries temporais, inclui como trabalhos futuros o estudo de outras abordagens como métodos estatísticos, os quais são também utilizados em algumas tarefas como predição e classificação. A implementação de algumas abordagens estudadas também está sendo realizada com intuito de avaliar esses métodos por meio de bases de dados sintéticas e reais, sendo as reais provindas da Usina de Itaipu.

Referências

- Chong, K. T. and Yoo, S. G. (2006). Neural network prediction model for a real-time data transmission. *Neural Computing & Application*, 15(3):373–382.
- Crone, S. (2002). Training artificial neural networks for time series prediction using asymmetric cost functions. In *9th International Conference on Neural Information Processing*, volume 5, pages 2374–2380.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *Second International Conference on Knowledge Discovery and Data Mining*, pages 82–88, Menlo Park, USA.
- Fu, A. W., Keogh, E., Lau, L. Y. H., and Ratanamahatana, C. A. (2005). Scaling and time warping in time series querying. In *31st International Conference on Very Large Data Bases*, pages 649–660.
- Geurts, P. (2001). Pattern extraction for time series classification. In *5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 115–127, London, UK. Springer-Verlag.
- Gupta, S., Ray, A., and Keller, E. (2005). Symbolic time series analysis of ultrasonic data for early detection of fatigue damage. *Mechanical Systems and Signal Processing*, pages 866–884.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, USA.
- Kadous, M. W. and Sammut, C. (2004). Constructive induction for classifying time series. *Lecture Notes in Computer Science*, pages 192–204.
- Keogh, E. and Kasetty, S. (2002). On the need for time series data mining benchmarks: A survey and empirical demonstration. In *8th International Conference on Knowledge Discovery and Data Mining*, pages 102–110, New York, NY, USA.
- Keogh, E. and Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. In *Knowledge and Information Systems*, volume 8, pages 154–177.

- Keogh, E., Lonardi, S., and Ratanamahatana, C. A. (2004). Towards parameter-free data mining. In *10th International conference on Knowledge Discovery and Data Mining*, pages 206–215, New York, New York, USA.
- Khatkhate, A., Ray, A., Keller, E., Gupta, S., and Chin, S. C. (2006). Symbolic time-series analysis for anomaly detection in mechanical systems. *IEEE/ASME Transactions On Mechatronics*, 11:439–447.
- Last, M., Kandel, A., and Bunke, H. (2004). *Data Mining in Time Series Databases*, volume 57 of *Machine perception and artificial intelligence*. Danvers, MA, USA.
- Last, M., Klein, Y., and Kandel, A. (2001). Knowledge discovery in time series databases. *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, 31(1):160–169.
- Latorre, M. R. D. O. and Cardoso, M. R. A. (2001). Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. *Revista Brasileira de Epidemiologia*, 4(3):145–152.
- Lendasse, A., V.Wertz, Simon, G., and M.Verleysen (2004). Fast bootstrap applied to ls-svm for long term prediction of time series. In *IEEE International Joint Conference on Neural Networks*, volume 1, pages 705–710, Budapest, Hungary.
- Liu, B. and Liu, J. (2002). Multivariate time series prediction via temporal classification. In *18th International Conference on Data Engineering*, page 268, Washington, DC, USA. IEEE Computer Society.
- Maimon, O., Kandel, A., and Last, M. (1999). Information—theoretic fuzzy approach to knowledge discovery in databases.
- Menescal, R. A., Vieira, V. P. P. B., Fontenele, A. S., and Oliveira, S. K. F. (2001). *Incertezas, ameaças e medidas preventivas nas fases de vida de uma barragem*. Seminário Nacional de Grandes Barragens, Ceará, Fortaleza, Brasil.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Boston, USA.
- Morettin, P. A. and Toloi, C. M. (2006). *Análise de Séries Temporais*. Edgard Blücher, São Paulo, Brasil, 2 edition.
- Nguyen, H. and Chan, W. (2004). Multiple neural networks for a long term time series forecast. *Neural Computing & Application*, 13(1):90–98.
- Pal, S. K. and Mitra, P. (2004). *Pattern Recognition Algorithms for Data Mining*. Chapman & Hall/CRC.
- Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., and Lendasse, A. (2005). Methodology for long-term prediction of time series. *Pascal Pattern Analysis, Statistical Modelling and Computational Learning*.
- Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining Knowledge Discovery*, 13(3):335–364.
- Wei, L. and Keogh, E. (2006). Semi-supervised time series classification. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 748–753, New York, New York, USA.