Uma Nova Abordagem Para Detecção de Outliers em Séries Temporais: Estudo de Caso em Consumo de Energia na Região Amazônica

Diemisom C. R. Melo, *Mestrando PPGEE, UFPA,* Adriana R. Castro, *Docente PPGEE, UFPA*

***Resumo*—A Detecção de Outliers é uma tarefa amplamente pesquisada na literatura com aplicação em diversas áreas de estudo. As Redes Neurais Artificiais Autoassociativas constituem uma das técnicas utilizadas para esta tarefa. Entretanto, não se encontra na literatura a aplicação destas redes em Detecção de Outliers em Séries Temporais. Dessa forma, este trabalho apresenta uma abordagem para a Detecção de Outliers em Séries Temporais, e propõe um estudo de caso realizado em um conjunto de dados de Consumo de Energia Elétrica na Amazônia brasileira, no Estado do Pará.**

***Palavras Chaves*-Detecção de Outliers, Redes Autoassociativas, Séries Temporais, Predição de Consumo de Energia.**

I. INTRODUÇÃO

SÉRIES temporais possibilitam o estudo de fenômenos complexos, e estão sujeitas a eventos inesperados ou mesmo incontroláveis. Esses eventos podem originar observações errôneas que de alguma forma são inconsistentes com as demais observações da série. Na literatura, estas observações são denominadas outliers, valores aberrantes, dados atípicos, observações discrepantes, entre outros. Barnett e Lewis propõem "Deve-se definir um outlier em um conjunto de dados como uma observação (ou subconjunto observações), que parece ser incompatível com o resto desse conjunto de dados" [1].

Outliers podem surgir por diferentes razões, seja devido a erros grosseiros ou a mudanças na série. Erros grosseiros são observações com defeito, como erros de medição, gravação e digitação. Eles devem ser naturalmente identificados e corrigidos sempre que possível devido seu potencial para alterar a correlação da série, influenciado resultados de previsão ou classificação, por exemplo.

II. TIPOS DE OUTLIERS EM SÉRIES TEMPORAIS

Os problemas mais comuns encontrados em séries temporais são caracterizados pela ausência de dados (presença de zeros ou nulos), mudança de nível e picos. Algumas vezes a ausência de dados possui significado, como por exemplo, uma falha no fornecimento de energia elétrica (blackout), mas em geral está associada a problemas de medição. Dependendo do tipo de estudo ou análise, a correção desses dados de medição torna-se necessária, uma vez que estatísticas básicas podem ser enviesadas e assim ocasionar conclusões equivocadas. Para estudos de previsão de séries temporais é importante que esses problemas sejam detectados e se possível corrigidos, pois a qualidade do histórico de dados afeta diretamente a qualidade da previsão [2].

A. Outliers em Séries Temporais

A **ausência de dados** (Figura 1.) pode afetar um subconjunto de observações e geralmente é visualmente detectado. Esta falha geralmente está associada a problemas no sistema aquisição ou medição de dados. Várias técnicas de substituição ou preenchimento podem ser empregadas e a aplicação de qualquer dessas técnicas depende de uma análise criteriosa para verificar a aderência dos resultados na série original.

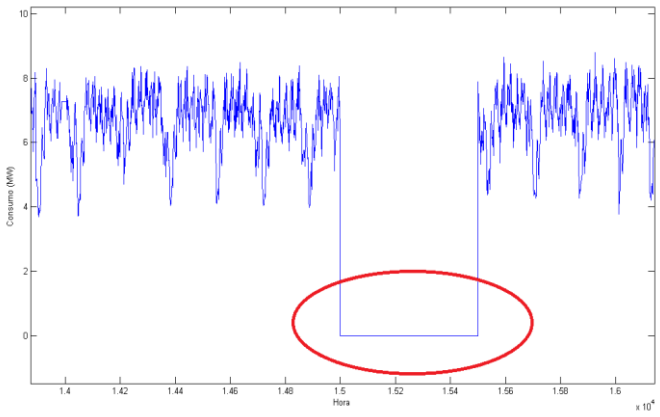

Figura 1 Dados Ausentes

A **mudança de nível** é identificada como uma mudança no valor médio da série e afeta um conjunto de observações contínuas (Figura 2.). Em muitos estudos esses dados não necessitam ser corrigidos, pois não são decorrentes de falhas de medição.

Os **picos** (Figura 3.), são observações que diferem do valor das demais observações no contexto que se apresenta, ou seja, são amostras cujo valor não seria considerado um outlier se ocorresse antes ou depois do momento que ocorreu.

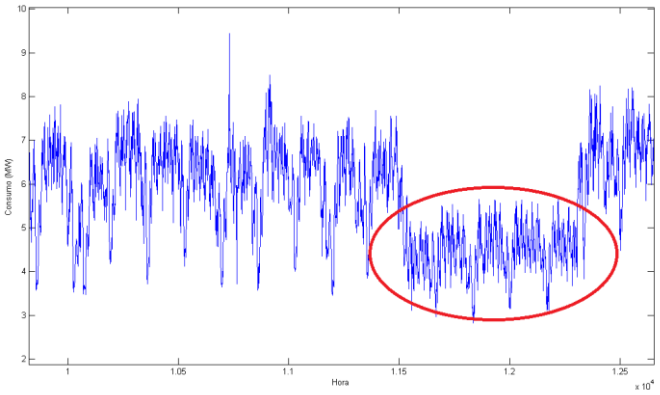

Figura 2 Mudança de Nível

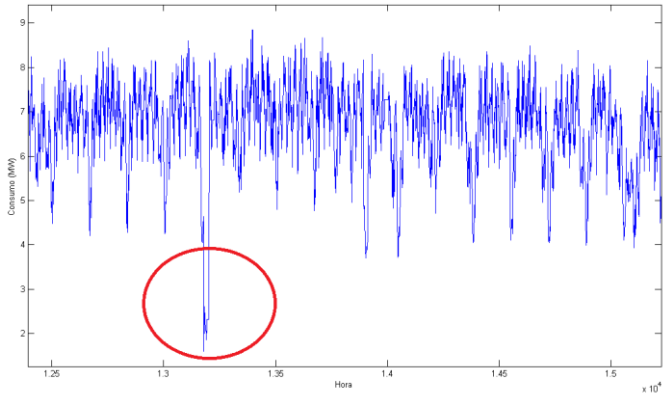

Figura 3 Pico

B. Tipos de Saída para o Detector de Outliers

Um dos requisitos para qualquer técnica de detecção de outlier é a maneira pela qual são relatados os outliers. Basicamente existem técnicas de rótulo e técnicas de pontuação.

Técnicas de Rótulo - atribuem um rótulo (normal ou outlier) para cada amostra, atuando como um algoritmo de classificação. O benefício dessas técnicas é que elas fornecem um conjunto exato de outliers para os analistas. A desvantagem é que elas não se diferenciam entre os diferentes valores de outliers, nenhum ranking de outliers é fornecido. Muitas vezes, deve existir um valor de confiança associado com um padrão ser ou não um outlier [2].

Técnicas de Pontuação - atribuem uma pontuação para cada padrão, dependendo do grau em que tal padrão é considerado um outlier. Assim a saída destas técnicas é uma lista ordenada de outliers. Um analista pode escolher se quer analisar os valores mais discrepantes ou usar um ponto de corte para selecionar os outliers. A desvantagem de uma lista ordenada de outliers é a escolha do limiar para selecionar um conjunto de valores discrepantes. Muitas vezes a escolha deste limite não é direta e deve ser fixada arbitrariamente [2].

III. Redes Neurais Autoassociativas Para Detecção de Outliers

As Redes neurais do tipo Perceptron Multicamadas (MLP-Multilayer Perceptron) já vêm sendo utilizadas para o desenvolvimento de sistemas de previsão de carga há um bom tempo. Uma MLP possui um número finito de camadas sucessivas (Figura 4.), cada uma tendo um número finito de unidades processadoras, chamadas neurônios [3].

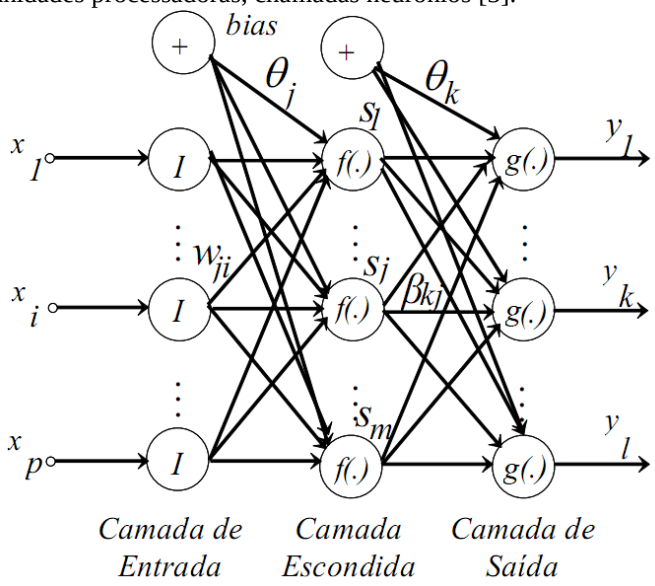

Figura 4 Perceptron de Múltiplas Camadas

Um tipo especial de redes neurais são as Redes Neurais Artificiais Autoassociativas (RNAA) [3]. Esta rede, mostrada na Figura 5, replica o vetor de entrada para a saída. Em outras palavras, a camada escondida da rede compacta as entradas e então as descompacta para recriar as entradas na camada de saída [3].

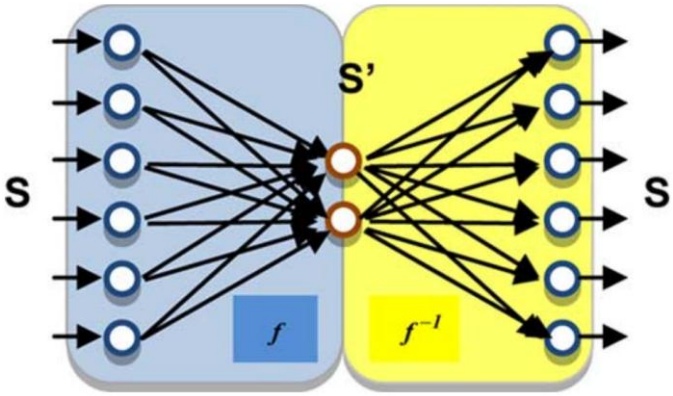

Figura 5 Rede Neural Artificial Autoassociativa

O erro entre o vetor de entrada e saída será baixo para qualquer sequência treinada. Quando um novo padrão contém dados ruins, as características deste padrão não pertencem às características aprendidas pela rede e o erro será elevado quando apresentado à RNAA. Portanto, diferenças anormais entre as entradas e saídas da RNAA indicam um padrão muito suspeito de conter dados ruins, ou seja, um possível outlier.

Quando algum outlier é apresentado à rede, ela falha em recriá-lo, então o outlier é detectado. A diferença entre a entrada e a saída da rede pode ser utilizada para medir o quanto uma amostra pode ser um outlier. Esta técnica opera de maneira semi supervisionada, assumindo a presença apenas de dados normais para o treinamento [3].

DOI: 10.5540/03.2013.001.01.0195

IV. Abordagem Proposta

Apesar do extenso uso de redes neurais artificiais para a tarefa de detecção de outliers, as redes autoassociativas não têm sido utilizadas para a detecção de outliers em séries temporais [2]. A abordagem deste trabalho consiste em utilizar as RNAA's para a detecção de outliers em séries temporais.

Para tanto, deve-se calcular o erro máximo de treinamento da rede, assumindo que o treinamento tenha sido executado apenas com dados normais. Este erro máximo será utilizado como um limiar entre dados normais e outliers, durante a fase de teste da rede. Os quatro passos que seguem descrevem detalhadamente a metodologia proposta:

1) Formar o novo conjunto de dados

Cada amostra desse novo conjunto de dados consiste em um vetor $X_n=\{x_n,x_{n-1},x_{n-2},x_{n-3},x_{n-4}\}$, em que x_n é o valor da série temporal no instante n; x_{n-1} é o valor da série temporal no instante n-1; x_{n-2} é o valor da série temporal no instante n-2; x_{n-3} é o valor da série temporal no instante n-3; e x_{n-4} é o valor da série temporal no instante n-4.

2) Treinamento da Rede Autoassociativa

Nesta etapa deve-se treinar a rede autoassociativa com o conjunto de dados gerado no passo 1.

3) Calcular o erro máximo de treinamento

Neste passo deve-se calcular o erro de treinamento da rede associado a cada amostra x_n. Dado que cada amostra original da série temporal é utilizada cinco vezes na série construída no passo 1, o erro de cada amostra corresponde a soma dos erros obtidos em cada aparição da amostra original no novo conjunto de dados. Uma vez calculado o erro de todas as amostras, o erro máximo de treinamento torna-se o limiar do sistema.

4) Testes

Para cada nova amostra obtida, deve-se formar um vetor X_n, fazendo uso dos valores anteriormente apresentados à rede. Em seguida, deve-se apresentar este novo padrão formado à rede autoassociativa, e acumular os erros para cada amostra da série temporal. Após a quinta utilização de uma amostra, pode-se calcular o erro total associado a ela e determinar se se trata de um outlier ou não.

V. Estudo de Caso em Série Temporal de Consumo de Energia

Como estudo de caso para testar a abordagem proposta foi utilizada um conjunto de dados de uma série temporal de consumo de energia elétrica no estado do Pará entre os anos de 2005 e 2006. O intervalo de medida do consumo foi de uma hora e a unidade de consumo é Mega Watts. A Figura 6 mostra a série temporal.

Utilizou-se 5 (cinco) neurônios nas camadas de entrada e de saída e 4 (quatro) neurônios na camada escondida. A RNA autoassociativa foi treinada utilizando a variante de Levenberg do algoritmo de retro propagação do erro (*Backpropagation*).

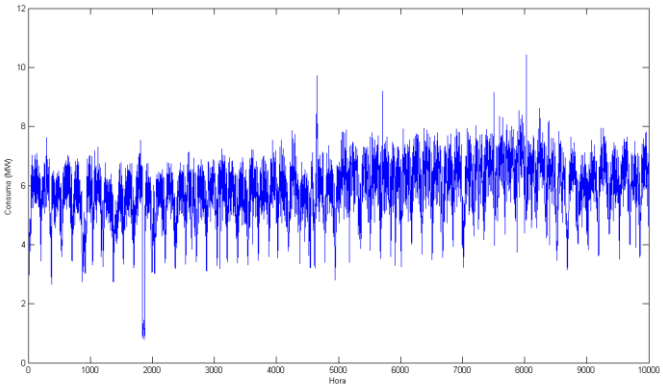

Figura 6 Série Temporal de Consumo de Energia

A série temporal continha 10 mil amostras e foi dividida em duas novas séries temporais, com 5 mil amostras cada, sendo a primeira para o treinamento da rede e a segunda para validação.

Vale ressaltar que durante o processo de execução dos passos que constituem a abordagem descrita anteriormente, o conjunto de dados gerado no passo 1 permite calcular o erro total apenas a partir da quinta amostra da série temporal original, entretanto, a amostra pode ser considerada um outlier a partir do momento que o valor do erro associado a ela seja maior que o limiar obtido no processo de treinamento da rede.

A. *Inserção de Outliers Virtuais*

Para viabilizar a análise do comportamento da abordagem proposta diante de outliers do tipo pico, foram alteradas quatro amostras presentes na série temporal de validação, conforme a Figura 7, além dos picos já existentes na série.

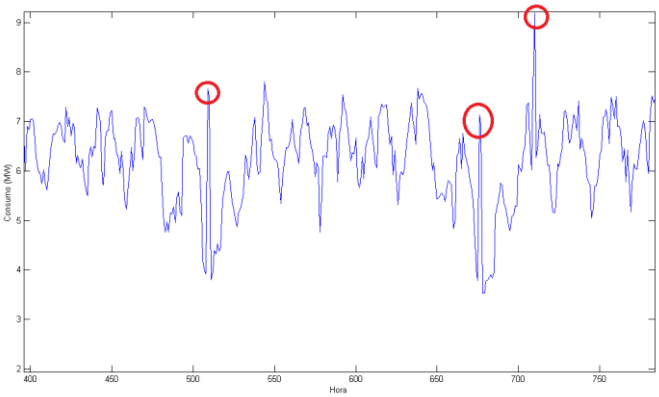

Figura 7 Picos Inseridos na Série de Validação

B. *Resultados Obtidos*

A Figura 8 mostra os erros de treinamento para cada amostra. O erro máximo obtido foi de 2.93, aproximadamente. Assim sendo, o limiar de outliers para o teste e validação da rede foi de 2.93. Valores acima desse passam a ser considerados outliers.

DOI: 10.5540/03.2013.001.01.0195

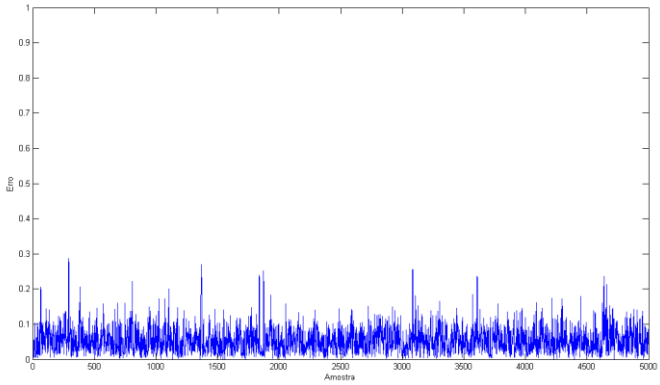

Figura 8 Erro de Treinamento

A Figura 9 mostra os erros obtidos para a cada amostra da série de validação. Valores circulados estão destacando o resultado alcançado para os outliers virtuais acrescentados. Além disso, a rede também reconheceu outros outliers na série temporal.

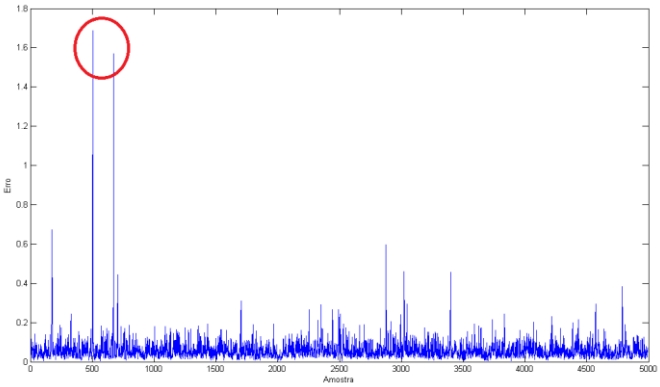

Figura 9 Erro de Validação

VI. CONCLUSÃO

Fica clara a capacidade de detecção de outliers das RNAs Autoassociativas aplicadas a séries temporais, segundo a abordagem proposta. Todos os outliers inseridos no conjunto de dados foram detectados pela rede. Pode-se ainda melhorar o uso da abordagem, alterando a saída do sistema de detecção, para que gere uma pontuação, representando o quanto uma amostra pode ser um outlier. Outra possibilidade de melhora é o uso do critério de entropia para otimização da rede neural. Pode-se também utilizar uma abordagem híbrida, com algoritmos genéticos, por exemplo, para proporcionar a correção de outliers. Um algoritmos genético poderia ser utilizado para minimizar o erro entre a entrada da rede e a saída obtida. Deve-se também adaptar a abordagem deste trabalho para séries temporais multivariadas.

REFERÊNCIAS

[1] Barnet, V. and Lewis, T., Outliers in Statistical Data, 3rd ed., New York.

[2] V. Chandola, "Anomaly Detection: A Survey," *ACM CSUR*, vol. 41, no. 3,Jul, 2009.

[3] M. Marko, S. Singh, "Novelty detection: a review—part 2: neural network based approaches" in ACM Signal Processing, vol. 83, no. 12, Dec, 2003;

