



DINCON 2017

CONFERÊNCIA BRASILEIRA DE DINÂMICA, CONTROLE E APLICAÇÕES

30 de outubro a 01 de novembro de 2017 – São José do Rio Preto/SP

Uma abordagem de uma rede multicamada para a estimação de agente contaminante no ar

Jean Pierre Brik López Vargas¹

Gilmar Barreto²

Faculdade de Engenharia Elétrica e Computação

Departamento de Semicondutores, Instrumentos e Fotônica

Laboratório de Controle e Sistemas Inteligentes, UNICAMP, Campinas, SP

Resumo. Usaremos uma rede neural multicamada para estimar a concentração de uma substância, benzeno (C_6H_6) encontrada no ar, a partir das medições de concentrações de outras três substâncias registradas por um dispositivo que mede a Qualidade do Ar. Os dados apresentam comportamento não linear, dados corrompidos, tendência e sazonalidade. O método que usamos para treinar a rede neural é de retro propagação e usamos validação cruzada para verificar a proximidade do resultado. Como resultado obtemos uma estimação bem aproximada da saída real. O propósito é mostrar com detalhe todos os passos para obter esses resultados, assim também as escolhas dos parâmetros como número de neurônios e épocas no processo de treinamento, assim mostrar o potencial das redes neurais para problemas complexos envolvendo sistemas do meio ambiente.

Palavras-chave. Rede Neural, Serie Temporal, Validação Cruzada, Qualidade do Ar.

1 Introdução

O meio ambiente é algo que devemos preservar assim nesse contexto é importante também analisar esse tipo de sistemas. A qualidade do ar é parte importante de nossa vida e é definida devido às diversas concentrações de substâncias contidas nele. Neste caso particular temos um dispositivo que recolhe dados de algumas substâncias que contaminam o ar. O nosso objetivo é estimar uma dessas substâncias a partir de outras, assim verificar o obtido pelo dispositivo e fazer algumas previsões dessa substância estimada. Considerando que as redes neurais são usadas cada vez mais em diversos campos como identificação, classificação, regressão, processamento, reconhecimento e para diagnóstico. Uma rede multicamada perceptrón(MLP), que é uma das mais conhecidas e usadas devido que têm

¹lopezvargasj@gmail.com, jean@dsif.fee.unicamp.br

²gbarreto@dsif.fee.unicamp.br

aproximação universal, esta rede têm a possibilidade de ter muitas camadas intermediárias as quais ajudam no processamento de grandes quantidades de dados. Esta rede é dividida em camadas de entradas que são multiplicadas por pesos sinápticos para depois fazer uma combinação linear desse resultado e ser avaliado numa função de ativação que geralmente no caso de MLP é uma função tangente hiperbólica a qual de acordo aos pesos se ajusta à saída desejada.

Uma representação do trabalho de estimação de uma rede multicamada.

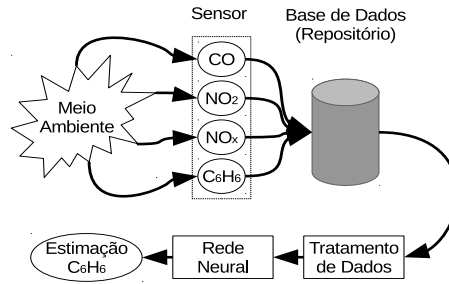


Figura 1: Representação em blocos

2 Métodos

A seguir descreveremos os passos que se consideraram para encontrar a estimação adequada do benzeno usando uma rede neural multicamada.

2.1 Dados de entrada e Saída

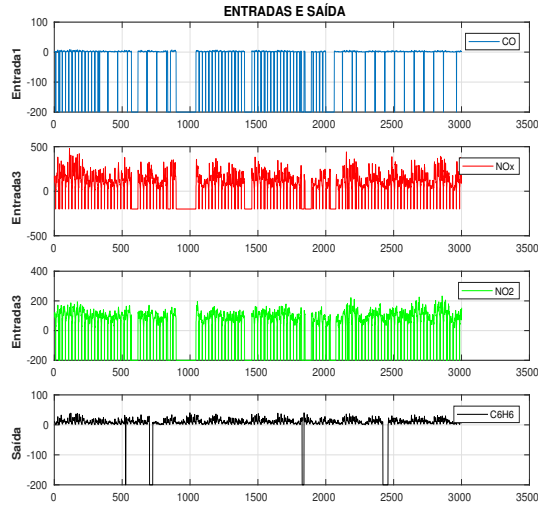
Usamos um repositório de Machine Learning [3]. Para esse propósito devemos examinar os dados que trabalharemos, assim nos escolheremos os dados mencionados em [5]. O número de entradas são três e de uma saída. Estes dados contém 9358 instâncias de resposta por hora em média, as quais são medidos por um arranjo de 5 sensores com capacidade de medir óxidos, químicos, metais, tudo isso embutidos num dispositivo. Estes dados foram registrados num ano, entre março de 2004 e abril de 2005 na cidade de Milão, Itália.

Esses dados da Qualidade do Ar, os sensores, a medição da concentração de benzeno, o qual é muito nocivo [4], é possível estimar através de outras três diferentes substâncias. O Monóxido de Carbono (CO) o qual é muito perigoso devido à toxicidade, os Óxidos de Nitrogênio (NO_x) que são poluentes da atmosfera e o Dióxido Nítrico (NO_2) que também é muito tóxico, essas medições são consideradas como entradas.

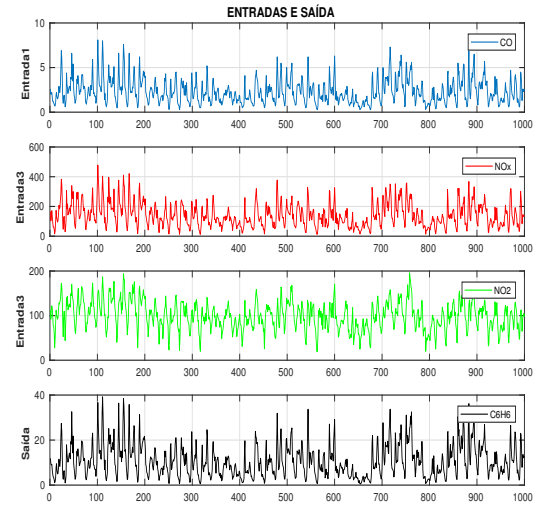
2.2 Análise dos dados

Analisando os dados observamos que os dados tens erros nas medições e quando o sensor tem falha ou uma medição errônea, ele registra um valor negativo grande (-200),

esto acontece devido a que quando as concentrações de CO, hidrocarbonetos não metano e outros são elevadas, se observam eventuais erros nos sensores. Nesse caso retiramos essas medições com erro do conjunto de dados, em geral como essas medições errôneas são aleatórias. Na figura (2a) a seguir, mostramos a representação dos dados de entradas e saída, os quais ainda não passaram pelo tratamento e na figura (2b) mostramos os dados filtrados que usaremos do repositório.



(a) Dados do repositório



(b) Dados analisados

Figura 2: Amostragem de dados

Depois de esse primeiro passo observamos também que os dados das diferentes entradas são de escalas diferentes, e sabemos que normalizando os dados geralmente se obtém uma convergência mais rápida, portanto aplicaremos essa técnica aos dados das três entradas e da saída.

Como resultado mostramos na figura (3) a diferença com o tratamento desses dados.

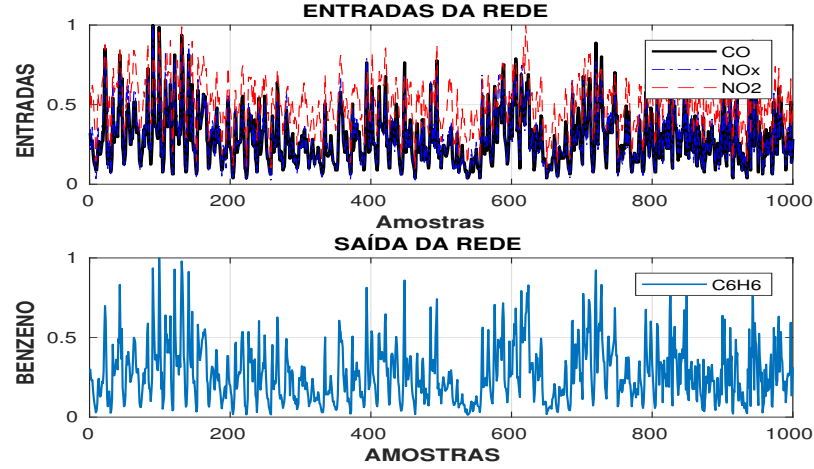


Figura 3: Dados entradas e saída

Após efetivar o tratamento nas medições também usaremos uma técnica de Validação Cruzada para separar em dois grupos os dados, o primeiro será o grupo com que o modelo trabalhará para obter os parâmetros o que geralmente é um 80% do total (814 dados), o segundo grupo do 20% restante (186 dados) se usará para validar que o modelo foi bem ajustando. Devemos considerar que a separação desses grupos deve ser realizada de forma que os dados de validação sejam representativos do total dos dados. A continuação separamos os grupos de treinamento (ajuste de parâmetros) e de validação (validar parâmetros).

2.3 Rede do MLP

A rede MLP usa um mapeamento não-linear de uma camada intermediária na qual é uma combinação linear das entradas com os pesos obtidos para depois usar funções de ativação, nós usaremos a função tangente hiperbólica. Para poder treinar apropriadamente a rede neural devemos definir algumas variáveis como o número de neurônios na camada intermediária e a quantidade de épocas que garanta a generalização da rede [1,6], essas variáveis serão definidas na subsecção a seguir.

2.4 Erro Quadrático Médio

Para obter o número de neurônios que devem ser usados na camada intermediária e as épocas necessárias para a generalização da rede neural usaremos o critério do erro médio quadrático. Neste caso para obter o número apropriado de neurônios a usar, comparamos e avaliamos o erro médio quadrático num intervalo de 1 a 50 neurônios.

Na figura (4), observamos que usando 30 neurônios na camada intermediária obtemos o erro quadrático médio (EQM) menor comparado com os outros EQM obtidos com o resto de neurônios, então decidimos trabalhar com 30 neurônios.

Uma vez encontrado o número adequado de neurônios treinamos a rede com a finalidade de estimar os pesos e a saída do sistema.

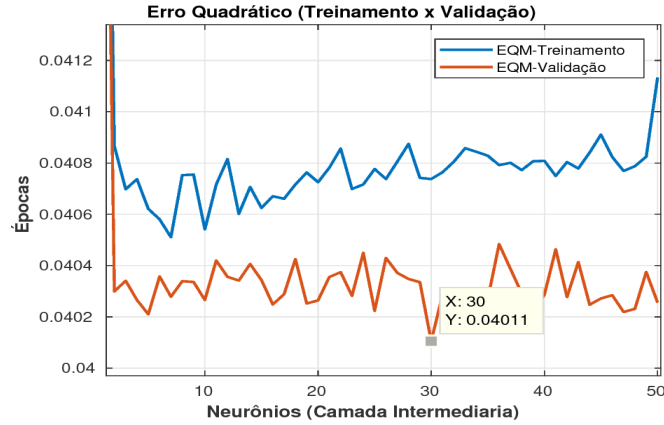


Figura 4: Erro quadrático médio para diferentes número de neurônios

3 Resultados

Conhecendo o número adequado de neurônios na camada intermediária da rede, precisamos conhecer o número de épocas adequadas para que a rede generalize bem os dados. Repetimos o mesmo critério usando o EQM, esta vez definindo como 30 o número de neurônios e considerando um número grande nas épocas, conseguindo assim sobre-treinar a rede.

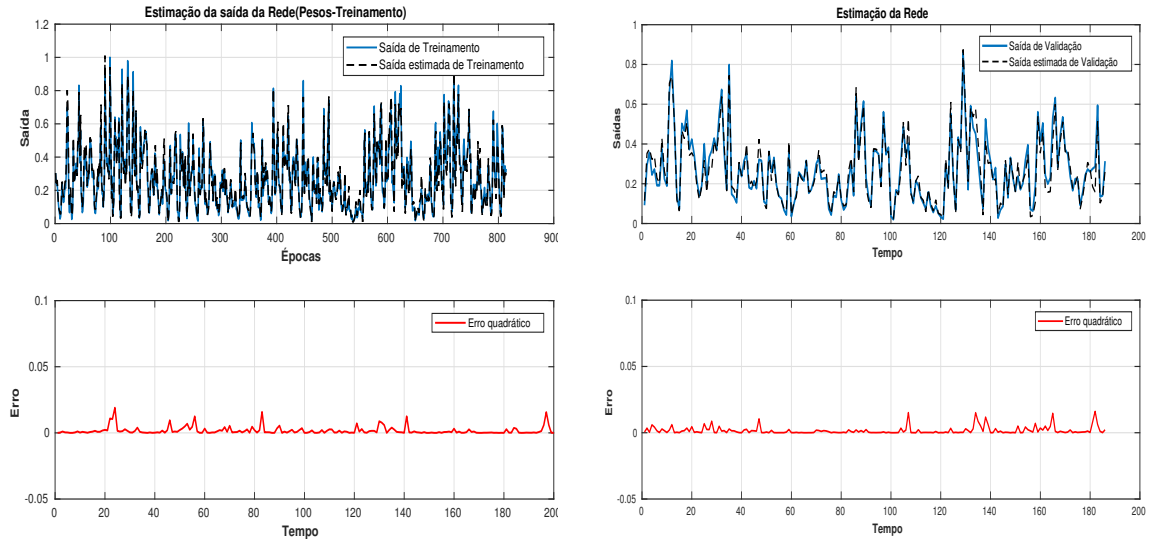
Na figura (5) observamos que na iteração 928 é adequado parar com o treinamento porque a partir desse ponto o EQM de validação se mantém e só sobre-treinaríamos a rede.



Figura 5: Erro quadrático médio para diferentes números de neurônios

Após o treinamento com o número adequado de neurônios e épocas, obtivemos como erro médio quadrático de (0.04). Na figura (6a), pode-se mostrar a estimativa da saída(conjunto de treinamento) obtida pela rede neural multicamada e a saída real e na figura (6b), mos-

tramos a estimativa da saída(conjunto de validação). Como podemos observar o treinamento da rede neural sobre o primeiro conjunto ocorreu de forma que está preparado para estimar novos dados do segundo conjunto.



(a) Saída(treinamento) estimada por MLP comparada com a saída real

(b) Saída estimada(validação) por MLP comparada com a saída real

Figura 6: Conjunto de Treinamento e Validação por uma Rede Multicamada

4 Conclusões

A partir dos resultados podemos concluir que com as concentrações das substâncias dos dados de entrada CO , NO_x e NO_2 podemos obter uma estimativa com baixo erro quadrático médio usando uma rede neural multicamada que generalize os dados de validação. Podemos perceber que as redes neurais podem ser usadas em diferentes áreas ajudando na resolução de problemas e realizando diferentes aplicações. Neste artigo vemos um clássico problema de estimação [2] aplicado à qualidade do ar, e podemos inferir que podemos estender esta ferramenta para problemas mais complexos e de grande relevância, devido as adversidades que a humanidade está atravessando com os câmbios climáticos e o aquecimento global. Com este artigo buscamos esclarecer o uso dessa rede convencional em conjuntos de dados com essas características de não linearidade, dificilmente estimadas por métodos de estimação clássicos.

Referências

- [1] HAYKIN, S. *Neural Networks and Learning Machines*. Pearson Education, 2011.

- [2] KOSKELA, T., LEHTOKANGAS, M., SAARINEN, J., AND KASKI, K. Time Series Prediction with Multilayer Perceptron , FIR and Elman Neural Networks. *In Proceedings of the World Congress on Neural Networks* (1996), 491–496.
- [3] LICHMAN, M. UCI machine learning repository, 2013.
- [4] ORWELL, R. L., WOOD, R. L., TARRAN, J., TORPY, F., AND BURCHETT, M. D. Removal of benzene by the indoor plant/substrate microcosm and implications for air quality. *Water, Air, and Soil Pollution* 157, 1 (Sep 2004), 193–207.
- [5] VITO, S. D., PIGA, M., MARTINOTTO, L., AND DI, G. Sensors and Actuators B : Chemical CO , NO 2 and NO x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. 182–191.
- [6] YAO, X. Evolving artificial neural networks, 1999.