Uso de técnicas de Data Mining na classificação de sorotipos da dengue

Jeferson Barros Alves

6 de novembro de 2015

Abstract. *The weather in Manaus, Amazonas is typically Equatorial warm and humid which causes a natural discomfort for its inhabitants. One of the main agents for the local weather regulation is the precipitations, whose prediction is very important although difficult, because the weather in the city is under the influence of many precipitation systems. This work aims at presenting an approach based on multilayered artificial neural networks to predict the occurrence of rainfall in Manaus. Data from an automatic meteorological station from* 1970 *to* 2014 *was used. After testing a set of* 2500 *different neural network that address this problem, it was possible to identify the network with* $4-9-7-1$ *architecture with best performance, being able to predict rainfall with almost* 70% *certainty.*

Resumo. *O clima na cidade de Manaus, Amazonas é caracterizado por calor e umidade tipicamente equatoriais, responsáveis por um desconforto natural a seus habitantes. Um dos principais agentes na regulação do clima local é a precipitação, sendo sua predição de apreciável importância, tratando-se porém, de um procedimento difícil, pois o clima da cidade é influenciado por diversos sistemas de precipitação. Este trabalho apresenta uma abordagem utilizando redes neurais artificiais de múltiplas camadas para prever a ocorrência de precipitações na cidade de Manaus. Foram utilizados dados climáticos de* 1970 *a* 2014*, provenientes de uma estação meteorológica automática. Após testar um conjunto de* 2500 *redes neurais que endereçam este problema, foi possível identificar a rede com arquitetura* $4-9-7-1$ *com melhor desempenho para este problema, capaz de prever precipitações com cerca de* 70% *de acerto.*

0.1. Introdução

Um dos fatores importantes para a sensação de conforto na cidade de Manaus, Amazonas é a ocorrência de precipitações, pois o clima nesta cidade é caracterizado pelo calor e pela umidade. As precipitações em Manaus ocorrem de maneira mais abundante entre os meses de Outubro a Julho e, segundo Sioli [Sioli 1991], a chuva em Manaus é abundante e não uniforme.

A ocorrência de precipitações em Manaus é resultado da influência de diversos sistemas de precipitação, a exemplo da Zona de Convergência Intertropical, Alta da Bolívia, Zona de Convergência do Atlântico Sul, dentre outros [Alavares et al. 2014, da Silva 2012]. Levando isto em consideração, é possível afirmar que a ocorrência de precipitações em Manaus é um fenômeno complexo e difícil de prever.

A previsão de precipitações em diversos lugares do mundo é um problema bastante endereçado pelas *Redes Neurais Artificiais* (RNAs). RNAs são sistemas paralelos distribuídos compostos por unidades de processamento simples (neurônios artificiais) que calculam determinadas funções matemáticas. Estas redes são muito utilizadas em problemas de classificação, categorização e previsão, tais como reconhecimento de imagens, detecção de fraudes, mineração de dados, dentre diversos outros [de Pádua Braga et al. 2007].

Este modelo de computação é bastante utilizado neste tipo de problema por várias razões, a citar: (i) RNAs são dirigidas por dados (*data driven*) e não demandam pré-requisitos restringentes sobre o que está sendo modelado; (ii) RNAs podem predizer

padrões que não são fornecidos durante o treinamento, isto é, são capazes de generalizar; (iii) RNAs são eficientes no treinamento de grandes amostras de dados graças a sua capacidade de processar em paralelo; (iv) RNAs possuem a habilidade de detectar relações complexas e não-lineares entre as variáveis dependentes e independentes [Darji et al. 2015].

Considerando a importância das RNAs no contexto de previsão de precipitações, este trabalho apresenta uma abordagem para previsão da ocorrência de precipitação na cidade de Manaus baseada na utilização de RNAs. Optou-se por adotar o modelo de *Redes Neurais Multicamadas Feedforward*, as quais foram treinadas com dados oriundos de cerca de 40 anos de registros climatológicos, que incluem temperatura máxima, mínima, umidade, etc. Foram construídas 2500 RNAs com arquiteturas diferentes e o objetivo era encontrar a rede com maior porcentagem de acertos na previsão de chuva. Como resultado, a melhor rede encontrada neste trabalho possui arquitetura $4 - 9 - 7 - 1$ e é capaz de prever precipitações com cerca de 70% de precisão.

Para apresentar estes resultados, o trabalho está organizado como segue. Uma breve fundamentação sobre RNAs é mostrada na Seção 0.2. A Metodologia adotada neste trabalho encontra-se descrita na Seção 0.3.3. Os resultados são apresentados e discutidos na Seção 0.4. Por fim, as considerações finais e trabalhos futuros encontram-se na Seção 0.5.

0.2. Redes Neurais Artificais

Inspiradas no cérebro humano, no qual neurônios realizam contínuas computações paralelas, as *Redes Neurais Artificais* (RNAs) são sistemas paralelos e distribuídos compostos por unidades de processamento estruturalmente simples e que computam determinadas funções matemáticas.

Para resolver um problema utilizando este modelo de computação, inicialmente são apresentados exemplos à rede. Durante esta fase, chamada *fase de treinamento*, a rede extrai características e padrões relevantes sobre a entrada. Posteriormente, durante as *fases de validação e testes*, estas informações serão utilizadas para obter respostas sobre o problema proposto [de Pádua Braga et al. 2007].

Um *neurônio artificial* é a unidade fundamental de processamento de informação em uma rede neural. O modelo de neurônio mais simples provê uma combinação linear de N pesos $w_1, \ldots, w_N$ e N entradas $x_1, \ldots, x_N$, cujo resultado é passado por uma função não-linear f, como mostrado na Figura 1.

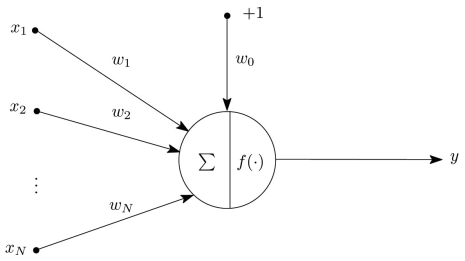

Figura 1: Representação de um neurônio artificial.

A entrada de um neurônio é um vetor $\mathbf{x} = [x_1, \ldots, x_N, 1]^T$, enquanto $\mathbf{w} = [w_1, \ldots, w_N, w_0]^T$ é um vetor de pesos do neurônio. O peso w_0 é o peso que corresponde ao viés da entrada, normalmente sendo igual a 1. O operador de soma realiza um mapeamento $\mathbb{R}^n \rightarrow \mathbb{R}$. A função $f : \mathbb{R} \rightarrow (0, 1)$, chamada *função de ativação*, é monótona,

contínua e sigmóide, na maioria dos casos [Mandic and Chambers 2001]. A saída y é obtida por meio da seguinte equação:

$$y = f\left(\sum_{i=1}^{N} w_i \cdot x_i + w_0\right). \tag{1}$$

Na computação neural, este tipo de neurônio segue o modelo proposto por McCulloch e Pitts [McCulloch and Pitts 1943]. Neurônios individuais, porém, possuem capacidade computacional limitada. Entretanto, quando conectados em uma rede – *rede neural* – são capazes de resolver problemas de mais alta complexidade [de Pádua Braga et al. 2007]. Modelos de redes neurais são especificados pela topologia da rede, características dos neurônios e regras de aprendizagem ou treinamento [Mandic and Chambers 2001].

Um dos modelos de redes neurais, as chamadas *Redes Neurais Multicamadas Feedforward* (RNMF), são caracterizadas pela presença de uma ou mais *camadas ocultas*, compostas por certa quantidade de neurônios, conforme ilustrado na Figura 2. O termo "oculto" refere-se ao fato de que estas camadas não são vistas diretamente a partir da entrada ou da saída da rede. A função destas camadas ocultas de neurônios é intervir entre a entrada externa e a saída da rede, permitindo a extração de estatísticas de alta-ordem [Haykin 2009].

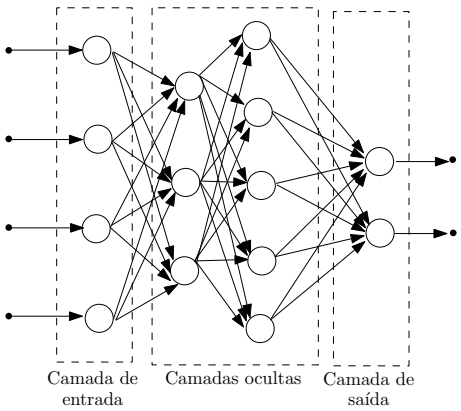

Figura 2: Exemplo de uma rede neural com múltiplas camadas e arquitetura $4-3-5-2$.

As RNAs se aplicam a problemas em que existem dados, experimentais ou gerados por meio de modelos, por meio dos quais a rede adaptará os seus pesos visando a execução de uma determinada tarefa. Há cinco grupo principais de tarefas às quais as RNAs se aplicam: classificação, categorização, aproximação, previsão e otimização. Considerando estas tarefas, há aplicações das RNAs em diversos setores e problemas, dentre os quais se destacam diagnóstico médico, detecção de fraudes, análise de expressão gênica, agrupamento de clientes, previsão financeira, previsão do tempo, dentre outros [de Pádua Braga et al. 2007].

0.3. Predição de Precipitação em Manaus, Amazonas

Manaus é a capital do estado do Amazonas no norte do Brasil e está localizada na região do baixo platô da planície do Rio Negro. Fundada por volta de 1693 hoje é a 7^{a} cidade mais populosa do país com cerca de 2 milhões de habitantes no ano de 2014 [Moreira and de Sene 2010].

De acordo com o sistema de classificação climática de *Köppen*, Manaus possui um clima tropical de monções, devido às altas temperaturas constantes durante todo o ano e as intensas chuvas de outubro a junho. Seu clima não categoriza-se como de floresta tropical devido ao mês mais seco do ano, agosto, o qual apresenta menos de 60 mm de precipitação [Alavares et al. 2014].

A região da cidade de Manaus e seus arredores possuem condições climáticas naturalmente desagradáveis, com calor e umidade característicos de regiões equatoriais, sendo ocasionados principalmente por dois fatores: sua localização geográfica e topografia [da Silva 2012]. Muitas de suas precipitações são resultados de vários sistemas de precipitação, como a zona de convergência Intertropical, Anticiclones Subtropicais, Alta da Bolívia, Sistemas de Mesoescala Tropical, Zona de Convergência do Atlântico Sul, Sistemas de Escala Sinóptica, dentre outros [da Silva 2012]. A chuva na cidade de Manaus é abundante, não uniforme e um importante fator que modela o clima da região [Sioli 1991], porém, constata-se que é um processo resultante de sistema natural de elevada complexidade.

0.3.1. Dados de Entrada e Saída

Os dados utilizados para a realização deste trabalho provêm de uma *estação meteorológica automática* composta pelos subsistemas de: (i) coleta de dados, cujo objetivo é integrar diversos sensores que aferem os parâmetros climáticos de interesse; (ii) armazenamento, que mantêm os dados em uma memória não volátil; (iii) comunicação, cuja tarefa é a transmissão dos dados da estação; e (iv) energia, responsável por manter a estação funcionando por meio de energia solar. Os dados produzidos pela estação são validados e então armazenados em um banco de dados disponibilizado pelo INMET (Instituto Nacional de Meteorologia) [INMET 2015a].

Figura 3: Estação Meteorológica Automática utilizada pelo INMET.

Considerando o problema da previsão de precipitação na cidade de Manaus, foi utilizada a estação meteorológica do INMET, localizada em $3°6'13.2552''$S, $60°0'6596''$W, a 61.25m acima do nível do mar [INMET 2015b]. Esta estação encontra-se ilustrada na Figura 3. Foram recuperados dados climáticos registrados diariamente entre 01 de Janeiro de 1970 a 31 de Dezembro de 2014, resultando num conjunto com cerca de 13179 dias de dados válidos. A partir destas observações, foram colhidos os seguintes parâmetros meteorológicos:

1. Temperatura Máxima Média;
2. Temperatura Mínima Média;
3. Umidade Relativa Média;
4. Velocidade Média do Vento;
5. Precipitação Média.

No contexto deste trabalho, os parâmetros de 1 a 4 serão utilizados como entradas das redes neurais, enquanto que o parâmetro 5 será a saída, ou seja, o valor que se deseja prever.

0.3.2. Modelo de Redes Neurais

O modelo de *Redes Neurais Multicamadas Feedforward* (RNMF) será utilizado para realizar a tarefa de predição proposta. Este modelo foi escolhido por ser capaz de aproximar, com precisão aceitável, uma função arbitrária [Barreto 2002].

A utilização do modelo RNMF em tarefas de previsão de chuva já mostra-se consolidada na literatura. Trabalhos como o de Abhishek et al. [Abhishek et al. 2012] e Luk et al. [Luk et al. 2001] mostram que RNMFs apresentam resultados similares quando comparadas a modelos de maior complexidade em problemas de predição de chuva. Estes resultados reforçam a escolha do modelo em questão no contexto deste trabalho.

No trabalho de Luk et al. [Luk et al. 2001], em particular, este modelo de redes neurais foi capaz de gerar alertas contra inundações com 15 minutos de antecedência. Já no trabalho de Abhishek et al. [Abhishek et al. 2012], estas redes foram utilizadas para prever monções na Índia. Ambos os trabalhos mostram aplicações práticas e relevantes das RNMFs.

0.3.3. Metodologia

Antes de apresentar os dados para o treinamento das redes neurais, um processo de *normalização pela média* foi realizado, conforme

$$x'_i = \frac{x_i - \mu}{\sigma}, \qquad \forall x_i \in X, \tag{2}$$

em que X é o conjunto de dados a ser normalizado, μ e σ são os valores da média aritmética e do desvio padrão do conjunto X, respectivamente. Isto ocorreu para reduzir a grande variabilidade dos dados, visando a melhor convergência do algoritmo de retropropagação de erros (*backpropagation*).

Os 13179 dias de dados válidos foram separados em três conjuntos: (i) *treinamento*, cujos dados apresentarão as características do problema à rede; (ii) *validação*, cujos dados servirão para avaliar o aprendizado da rede; e (iii) *teste* do qual os dados são destinados a averiguar a capacidade de generalização da rede. A separação destes conjuntos visou segmentar de maneira temporal o conjunto de dados de entrada. Desta forma, o conjunto de treinamento recebeu dados entre 1970 e 2010 (86% dos dados); ao conjunto validação foram assinalados os anos de 2011 a 2013 (cerca de 11%); enquanto que ao conjunto teste foram designados 365 dias de dados (aproximadamente 3%), relativos ao ano de 2014. Tal separação foi feita com o objetivo de agrupar características de memória inerentes aos processos climáticos, ao mesmo tempo que tenta-se evitar o *overfiting*, uma vez que cada dado do conjunto será apresentado à rede de forma randômica.

Embora redes neurais com somente uma camada escondida possam aproximar qualquer função contínua, optou-se por utilizar redes com duas camadas escondidas na expectativa de que estas pudessem capturar melhor os detalhes envolvidos no mapeamento de uma função de predição. Para isso, foram treinadas redes com $X = 4$ neurônios na camada de entrada, Y e Z neurônios na primeira e segunda camadas escondidas, respectivamente, em que $1 \leq Y, Z \leq 50$. Há apenas 1 neurônio na camada de saída, correspondendo a previsão realizada pela rede.

Um total de 2500 redes neurais diferentes poderiam satisfazer os requisitos previamente apresentados. Para automatizar a tarefa de treinamento, validação e testes dessas redes, bem como para a posterior análise dos dados, o software MATLAB® foi utilizado.

Dado o amplo número de redes neurais a considerar, a escolha das arquiteturas mais apropriadas foi efetuada em duas etapas:

1. **Treinamento**. Considerou o *Erro Médio Quadrático* (MSE - *Mean Squared Error*) da previsão de chuva como parâmetro guia para a avaliação do aprendizado;
2. **Avaliação**. Nesta etapa, as redes foram analisadas tendo como medida de performance a porcentagem de acertos na fase de testes.

O MSE não poderá ser usado como métrica confiável para a escolha das redes, pois, por tratar-se de um trabalho de classificação, suas saídas fornecerão a predição da probabilidade de ocorrer precipitação em um determinado dia. Mesmo em problemas de regressão um baixo MSE pode ser uma medida de performance errônea, principalmente nos casos onde ocorre *overfiting*. Por este motivo, as arquiteturas de rede selecionadas serão as que possuírem maior porcentagem de acertos na fase de testes e que, por conseguinte, efetuaram a tarefa de predição mais precisamente.

As redes foram treinadas utilizando a otimização de *Levenberg-Marquardt* que é uma aproximação do método de Newton e que utiliza uma taxa de aprendizagem variável [Fun and Hagan 1996]. Geralmente é o algoritmo *backpropagation* mais rápido, embora requeira mais memória que outros algoritmos disponíveis [Mathworks 2015]. Utilizou-se a função tangente sigmoidal como função de ativação dos neurônios da rede, sendo esta muito utilizada em problemas de classificação como o abordado neste trabalho [Survey].

Os parâmetros de treinamento da rede foram mantidos fixos e encontram-se listados na Tabela 1, sendo a diferenciação entre as diversas redes testadas dada pela variação do número de neurônios das camadas escondidas.

Tabela 1: Parâmetros utilizados no treinamento das redes neurais.

Parâmetro	Valor
erro desejado	0.5×10^{-4}
épocas de treinamento	1000
erros máximos	7
gradiente mínimo	1×10^{-10}
mu	0.003
decremento do mu	0.01
incremento do mu	7
mu máximo	1×10^{10}
tempo de treinamento	indeterminado

Utilizando a metodologia proposta, dentre as 2500 redes neurais consideradas, aquelas com melhor performance para previsão de chuva em Manaus foram identificadas. Os resultados obtidos, suas descrições e análises são apresentadas na seção a seguir.

0.4. Resultados e Discussões

Foi realizada uma busca exaustiva no conjunto das possíveis arquiteturas de redes neurais a serem utilizadas como um modelo de predição. Uma vez que o treinamento fora

concluído, os resultados produzidos foram avaliados resultando na seleção de 3 arquiteturas com melhor desempenho, no contexto das métricas de avaliação adotadas, cujas informações encontram-se na Tabela 2.

Tabela 2: Arquiteturas de redes selecionadas e seus parâmetros de avaliação

Arquitetura	**MSE**	**Acertos** $(\%)$
$9-7$	0.1930	67.67
$23-13$	0.1870	70.14
$43-20$	0.2069	66.30

A porcentagem de acertos, apesar de se mostrar levemente semelhante entre as diversas arquiteturas verificadas, possui maior variabilidade, com desvio padrão de 0.0367, fornecendo um melhor *insight* a respeito do desempenho da arquitetura da rede. A comparação de performance das arquiteturas selecionadas baseadas neste parâmetro encontra-se na Figura 4. Nela, podemos constatar que mesmo um aumento substancial de neurônios nas duas camadas escondidas pode não produzir melhora significativa na predição das redes, como no caso da terceira arquitetura.

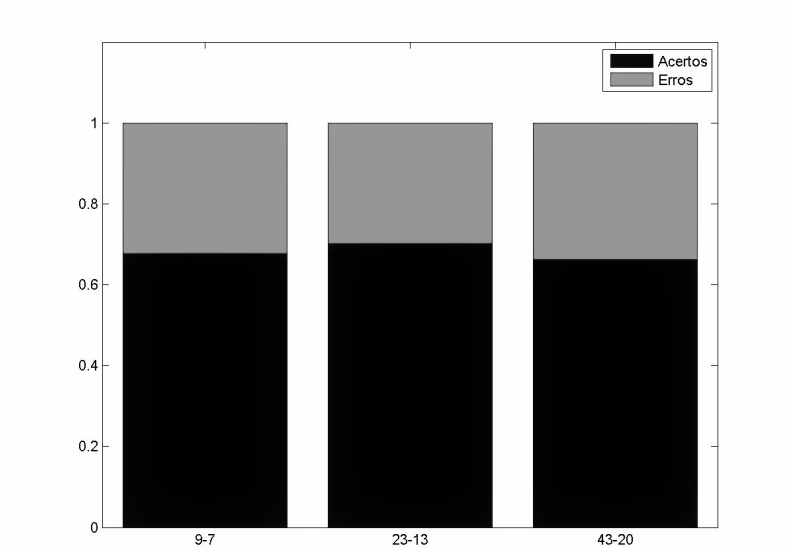

Figura 4: Comparação entre do desempenho das três redes selecionadas.

A fim de definir a melhor arquitetura, dentro dos parâmetros avaliados neste trabalho, seria necessário estimar, de forma mais confiável, os resultados produzidos pelas redes neurais selecionadas. Para isso, foram gerados os intervalos de confiança de cada arquitetura, que executaram 10 previsões correspondentes aos dias do ano 2014, das quais foram salvas as porcentagens de acerto, como amostras. Uma vez que o nível de significância empregado foi de $\alpha = 0.05$, pode-se afirmar que fora obtido, de forma fundamentada, limites de confiança que conterão o valor de acerto médio das redes neurais com uma certeza de 95%.

Analisando a Figura 5, na qual os intervalos de confiança das três redes encontram-se denotados, nota-se que as médias inferidas dos percentuais de acerto de cada arquitetura de rede podem englobar-se dentro dos intervalos de confiança umas das outras, o

que demostra certa equivalência entre o desempenho das redes. Por esta razão, é possível afirmar que a arquitetura de rede neural mais adequada para a tarefa de predição proposta é a que contêm 9 e 7 neurônios em sua primeira e segunda camada intermediária, respectivamente. Esta rede foi escolhida por possuir menor número de neurônios que as outras redes, o que ocasiona menor esforço computacional, enquanto apresenta considerável capacidade de generalização.

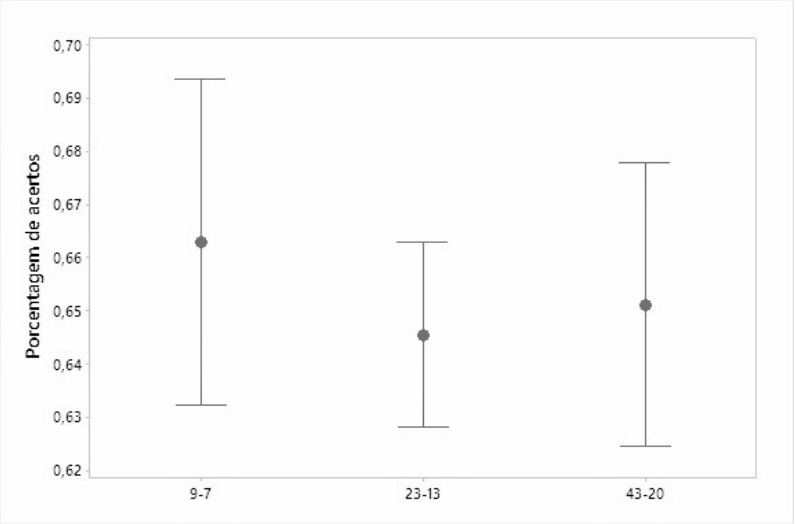

Figura 5: Intervalos de confiança para as arquiteturas de rede selecionadas. No eixo X, a indicação da arquitetura da rede; no eixo Y o percentual de acertos da rede.

A rede apresentou boa convergência em seu treinamento não tendo muitas flutuações. Uma demonstração da previsão de chuva é dada na Figura 6, onde pode-se observar a relação entre os dados de precipitação observados no ano de 2014 e os dados previstos pela rede identificada no escopo deste trabalho.

0.5. Considerações Finais

Neste trabalho foi ilustrada a aplicação de redes neurais multicamadas *feedforward* ao problema de previsão de precipitações na cidade de Manaus, Amazonas. Para endereçar este problema, foram utilizados dados obtidos de uma estação meteorológica localizada na cidade em questão durante os anos de 1970 a 2014. Como entrada, considerou-se os dados diários de temperatura máxima, temperatura mínima, umidade relativa e velocidade do vento. Como saída, considerou-se a ocorrência ou não de precipitação no dia em questão. Os dados foram separados em três conjuntos, sendo o ano de 2014 utilizado para os testes dos dados.

Considerando este cenário, foram consideradas arquiteturas de redes neurais com 2 camadas escondidas e com até 50 nós em cada uma dessas camadas, resultando em um total de 2500 redes. Os dados foram apresentados à estas redes e três delas foram classificadas como tendo melhor desempenho.

Dentre as três redes identificadas, os intervalos de confiança para o número de acertos das mesmas foi construído e observou-se que, ao nível de significância de 95%, elas são equivalentes em termos de desempenho. Assim, selecionou-se a rede com arquitetura $4-9-7-1$ como sendo a melhor dentre as três por possuir menos neurônios. Os resultados obtidos neste trabalho mostraram que, com a utilização de redes neurais, foi possível prever chuvas em Manaus com até 70% de acerto.

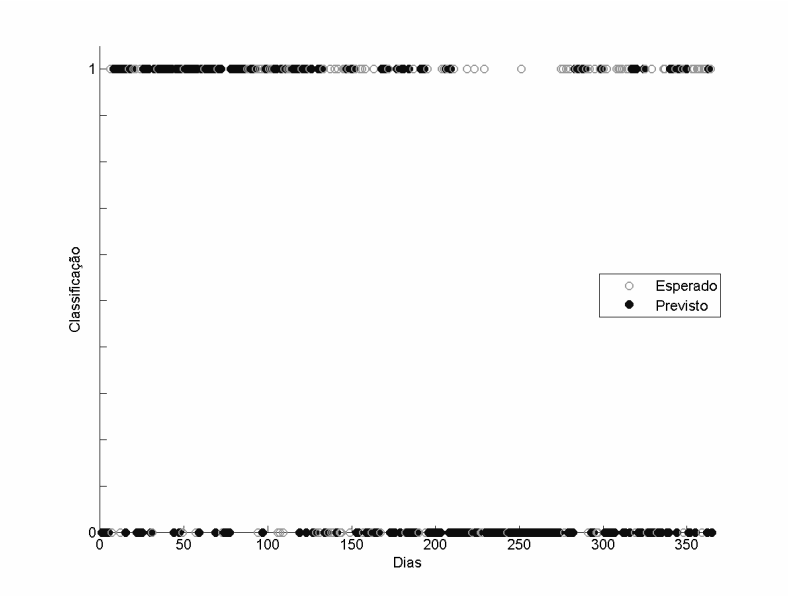

Figura 6: Resultados observados versus previsto da rede $4-9-7-1$ para o ano de 2014. No eixo X os dias observados, enquanto que no eixo Y, temos a classificação dos dias de dados, onde $Y=1$ indica presença de precipitação e $Y=0$ ausência.

Em trabalhos futuros, almeja-se aumentar a porcentagem de acertos das redes. Para tanto, espera-se considerar a recorrência e a relação temporal entre os dados de entrada. Se possível e disponível, deseja-se considerar mais variáveis climatológicas como entrada. Além disso, a previsão antecipada de chuva em Manaus é algo a ser investigado, considerando não só as redes neurais, como também outros modelos que posssam endereçar este tipo de situação.

Agradecimentos

O autor Patrick Magalhães agradece o apoio financeiro provido pela Samsung para a realização deste trabalho. Os autores agradecem a Profa. Maria Betânia Leal pelas sugestões fornecidas e que contribuíram para a melhoria deste trabalho.

Referências

Abhishek, K., M.P.Singh, Ghosh, S., and Anand, A. (2012). Weather forecasting model using artificial neural network. *Procedia Technology*, 4:311–318.

Alavares, C. A., Stape, J. L., Sentelhas, P. C., Gonçalves, J. L. M., and Sparovek, G. (2014). Koppen's climate classification map for brazil. *Meteorologische Zeitschrift*, 22(6):711–728.

Barreto, J. M. (2002). Introdução às redes neurais artificiais. UFSC - Universidade Federal de Santa Catarina.

da Silva, D. A. (2012). Função da precipitação no conforto do clima urbano da cidade de manaus. *Revista Geonorte*, 1(5):22–40.

Darji, M. P., Dabhi, V. K., and Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey. In *International Conference on Advances in Computer Engineering and Applications*, India. IMS Engineering College.

de Pádua Braga, A., de Leon F. de Carvalho, A. P., and Ludermir, T. B. (2007). *Redes Neurais Artificiais: Teoria e Aplicações*. LTC, 2 edition.

Fun, M. H. and Hagan, M. T. (1996). Levenberg-marquardt training for modular networks. *IEEE International Conference on Neural Networks*, 1:468–473.

Haykin, S. (2009). *Neural Networks and Learning Machines*. 3 edition.

INMET (2015a). Instituto nacional de meteorologia. `http://www.inmet.gov.br/portal/css/content/topo_iframe/pdf/Nota_Tecnica-Rede_estacoes_INMET.pdf`.

INMET (2015b). Instituto nacional de meteorologia. `http://www.inmet.gov.br`.

Luk, K. C., Ball, J. E., and Sharma, A. (2001). An application of artificial neural networks for rainfall forecasting. *Mathematical and Computing Modelling*, 33:683–693.

Mandic, D. P. and Chambers, J. A. (2001). *Recurrent Neural Networks for Prediction*. Wiley.

Mathworks (2015). Levenberg-marquardt backpropagation. `http://www.mathworks.com/help/nnet/ref/trainlm.html`.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133.

Moreira, J. C. and de Sene, E. (2010). *Geografia Geral e do Brasil*. Editora Scipione.

Sioli, H. (1991). *Amazônia: Fundamentos da ecologia da maior região de florestas tropicais*. Editora Vozes.