

Sistema Inteligente para Apoio ao Diagnóstico de Diabetes Empregando Redes Neurais

Maik Basso, João Paulo Vieira, Fábio José Parreira, Sidnei Renato Silveira,
Adriana Sadowski de Souza

maik@maikbasso.com.br, vieira.jpvieira@gmail.com, fabiojparreira@gmail.com,
sidneirenato.silveira@gmail.com, adrianasadowski@gmail.com

Universidade Federal de Santa Maria (UFSM) - Centro de Educação Superior Norte
(CESNORS) - Caixa Postal 54 - Frederico Westphalen - RS - Brasil
Departamento de Tecnologia da Informação

Resumo. Este artigo aborda o desenvolvimento de um protótipo de aplicação que utiliza técnicas de Inteligência Artificial, para apoiar o diagnóstico do Diabetes. O sistema apoia o diagnóstico, por meio de Redes Neurais Artificiais, do estado patológico do paciente, cujas informações para treinamento e teste do sistema foram extraídas da base de dados da comunidade indígena Pima. Comparando-se os resultados do sistema como as informações da base dados, foi obtida uma taxa de acerto de 81,31%.

Abstract. This paper presents the development of an application prototype, using Artificial Intelligence techniques, to support diabetes diagnosis. The system supports the patient pathological state diagnosis through Artificial Neural Networks, whose information for training and testing the system were extracted from the Pina Indian community database. Comparing the results of the system as the information of the database, an accuracy rate of 81,31% was obtained.

1. Introdução

Nos últimos anos a incidência do diabetes tem aumentado rapidamente em nível mundial. Recentemente, a Organização Mundial de Saúde (OMS) reconheceu que a doença é epidêmica. As estatísticas apontam que o número de casos, em todo o mundo, atualmente chega a 246 milhões. Até 2025, esse número deve chegar a 350 milhões, de acordo com a Federação Internacional de Diabetes (IDF). No Brasil, segundo dados do Ministério da Saúde, estima-se que existam cerca de 11 milhões de portadores de diabetes, sendo que 7,5 milhões já sabem que tem a doença (Oliveira; Vencio, 2014).

Alguns dos fatores que favorecem esse crescimento alarmante do diabetes são a obesidade, hereditariedade, hábitos alimentares inadequados e o sedentarismo. Estudos demonstram que 46,6% dos pacientes diagnosticados desconheciam o fato de serem portadores da doença, e dos pacientes sabidamente diabéticos, 22,3% não faziam nenhum tipo de tratamento (Botelho, 2003).

Embora a detecção de diabetes esteja melhorando, o tempo para diagnosticá-lo pode ser superior a 10 anos, a contar do início da doença, até a concretização do diagnóstico. Para diagnosticar um paciente com diabetes, o médico tem que analisar diversos fatores e isto dificulta o seu trabalho. Para auxiliar estes profissionais, da área de saúde, são utilizados sistemas computacionais com a finalidade de apoiar o diagnóstico. Tais sistemas processam informações, com mais detalhe e em menor tempo, quando comparados com aos seres humanos, proporcionando uma melhora na

qualidade dos serviços médicos, além de contribuir para a difusão de conhecimentos especializados (Kayaer; Yildirim, 2003).

Neste sentido, o presente trabalho apresenta um protótipo de sistema inteligente, empregando Redes Neurais multicamadas, para apoiar o diagnóstico de diabetes - Sistema Neural para Apoio ao Diagnóstico de Diabetes (SND). O artigo apresenta um breve referencial teórico sobre as áreas envolvidas (diabetes, redes neurais e o banco de dados para treinar e testar o sistema), situa o estado da arte e apresenta a solução implementada.

2. Referencial Teórico

2.1. Diabetes

Uma pessoa diagnosticada com diabetes sofre de hiperglicemia, que por definição é o excesso de açúcar no sangue. Este excesso ocorre porque o pâncreas, órgão responsável pela produção da insulina, não consegue produzir a quantidade necessária deste hormônio para que a glicose possa entrar na célula, onde seria consumida, o que resulta no acúmulo desta no sangue. O diabetes não afeta só a capacidade do organismo de consumir açúcar, mas também a capacidade de utilizar outras fontes de energia como as proteínas e as gorduras (Baier, Hanson, 2004). Basicamente, existem três tipos principais de diabetes: o diabetes tipo 1, que afeta principalmente os jovens, o diabetes tipo 2, que, geralmente, afeta as pessoas adultas e idosas e o diabetes gestacional, que pode ocorrer durante a gestação (Mazzaferri, 1988).

Para o diabetes tipo 1 existem diversas denominações: diabetes insulino dependente, diabetes infanto-juvenil ou diabetes imunomediado. A característica marcante deste tipo de diabetes é a produção insuficiente de insulina, pois suas células sofrem o que se chama de destruição autoimune. Neste tipo de diabetes a produção de insulina do pâncreas é insuficiente, pois suas células sofrem o que se chama de destruição autoimune. Os portadores de diabetes tipo 1 necessitam de injeções diárias de insulina para manterem a glicose no sangue em valores normais. Embora este tipo de diabetes ocorra em qualquer idade, é mais comum em crianças, adolescentes ou adultos jovens (Oliveira; Vencio, 2014).

Já o diabetes tipo 2, denominado também de diabetes não insulino dependente ou diabetes do adulto, corresponde a 90% dos casos de diabetes e geralmente ele é detectado em pessoas obesas, com mais de 40 anos de idade. Atualmente este quadro vem sofrendo mudanças, pois há grande ocorrência em jovens, em virtude de maus hábitos alimentares, sedentarismo e e e da vida urbana. Neste tipo de diabetes, embora exista a presença de insulina, a sua ação é dificultada pela obesidade, o que é conhecido como resistência insulínica. Por ser pouco sintomático, este tipo, na maioria das vezes, permanece por muitos anos sem diagnóstico e sem tratamento o que favorece a ocorrência de suas complicações no coração e no cérebro (Botelho, 2003).

Por fim, o diabetes Gestacional, ocorre quando há a presença de glicose elevada no sangue, durante a gravidez. Ao final da gravidez, após o parto, geralmente a glicose no sangue, se normaliza. No entanto, as mulheres que apresentam ou apresentaram diabetes gestacional, possuem maior risco de desenvolverem diabetes tipo 2 tardiamente, o mesmo ocorrendo com os seus filhos.

2.2. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são um paradigma de processamento de informação inspirado no sistema neural biológico. As RNAs são sistemas massivamente paralelos e distribuídos, formados por unidades de processamento simples, que calculam determinadas funções, normalmente não lineares. Essas unidades são distribuídas em camadas, sendo interligadas por conexões, as quais se associam a pesos. Estes pesos armazenam o conhecimento representado na rede, servindo para ponderar as entradas recebidas por unidade constituinte (Haykin, 2001).

Uma das características mais importantes de uma RNA é a capacidade de aprender através de exemplos, denominados de conjunto de treinamento, sendo estes apresentados a ela durante a etapa de treinamento. Durante o aprendizado a rede consegue extrair os padrões existentes neste conjunto de dados e, após finalizar a etapa de treinamento, ela aprende o suficiente para generalizar o problema. Ao final, na etapa de teste, a RNA consegue produzir respostas satisfatórias para dados desconhecidos, que fazem parte do mesmo problema, para o qual ela foi treinada.

Segundo Palma Neto e Nicoletti (2005), os algoritmos de treinamento das RNAs podem ser agrupados em supervisionados e não supervisionados. Nos supervisionados, além das informações, cada exemplo de treinamento traz a classe a qual ele pertence. Nos grupos não supervisionados a RNA é treinada sem considerar as informações da classe a qual cada exemplo está associado.

Existem várias arquiteturas para a implementação de uma RNA, mas estas podem ser divididas, basicamente, quanto ao número de camadas, e pelos tipos de conexões entre os neurônios (Fausett, 1994). De acordo com o número de camadas, a arquitetura de uma RNA pode ser denominada de camada única ou de múltiplas camadas. As RNAs com múltiplas camadas distinguem-se de redes com camada única pela presença de camadas ocultas. Além disso, são utilizadas em problemas mais complexos, que necessitam analisar várias combinações das entradas da rede, como por exemplo, o diagnóstico de diabetes.

Quanto aos tipos de conexões, que definem a direção do fluxo de ativações, as RNAs são definidas como *feedforward* e *feedback*. As RNAs *feedforward*, também conhecidas como acíclicas ou não recorrentes, são estruturadas em camadas, e não possuem realimentação de suas saídas para as entradas. Estas podem ter de uma ou mais camadas em sua configuração. Dentre os algoritmos para treinamento de redes *feedforward*, o *back-propagation* está entre os que proporcionam uma maior eficiência. Já as RNAs *feedback*, também denominadas de cíclicas ou recorrentes, diferenciam-se das não recorrentes por terem em sua estrutura a realimentação das saídas para as entradas (Haykin, 2001).

Para que uma rede neural tenha sucesso na resolução de problemas práticos, é necessário que ela consiga representar funções lineares e não lineares, assim como o algoritmo de treinamento deve ser suficientemente robusto e genérico, não restringindo a eficácia na resolução de problemas. As redes neurais de múltiplas camadas, com o algoritmo para treinamento *back-propagation*, contemplam essas características (Massad et. al, 2004).

2.3 Base de dados Pima Indians

Para treinar e testar a aplicação da RNA no protótipo desenvolvido, foi utilizada a base de dados da Universidade da Califórnia (UCI, 2014), denominada *Pima Indians*

Diabetes (PID). As informações para compor esta base foram coletadas na comunidade indígena *Pima*, que vive perto de Phoenix, Arizona, Estados Unidos. Todas as pessoas, cujos dados foram coletados, são mulheres, com idade igual ou superior a 21 anos.

Diversas pesquisas foram realizadas nesta comunidade, pois ela é conhecida por apresentar a maior taxa de incidência de diabetes do mundo. Entre os adultos 50% são diabéticos, além de apresentarem uma alta prevalência de obesidade (Baier e Hanson, 2004). Devido a estas características, os pesquisadores resolveram criar a base de dados PID, que é composta por 768 pessoas, onde 268 destas foram classificadas como tendo diabetes, do tipo 2, o que resulta em 34,9% da base. Os outros 65,1% não possuem diabetes e são classificados como normais.

Para cada pessoa que compõe a base, é informado: o número de vezes que a paciente ficou grávida, teste oral de tolerância à glicose, pressão arterial diastólica (mm Hg), espessura da dobra cutânea do tríceps (mm), idade (anos), índice de massa corporal (IMC) e a classificação (possui ou não diabetes).

3. Estado da Arte

Existem diversos estudos focados no desenvolvimento de mecanismos de diagnóstico de diabetes. Nas últimas décadas, as abordagens utilizadas extrapolaram o escopo clássico, voltado para ferramentas matemáticas, adotando técnicas inteligentes, com o intuito de aprimorar a flexibilidade do mecanismo em relação à dinâmica apresentada pelo processo de diagnóstico. Dentre as várias abordagens utilizadas, as redes neurais artificiais vêm sendo largamente utilizadas para auxiliar no diagnóstico de diabetes.

Polat (Polat et. al., 2008) apresenta uma abordagem, para diagnóstico de diabetes tipo 2, usando análise de componente principal (PCA) com sistema adaptativo neuro-fuzzy. Inicialmente, o PCA foi utilizado para reduzir as informações da base de dados, passando de 8 para 4. No segundo estágio, as 4 informações resultantes do PCA são fornecidas ao sistema classificador adaptativo neuro-fuzzy, que as classificam como sendo diabéticas ou normais. Este trabalho obteve, como resultado final, uma precisão de 82,05%, utilizando a base de dados *Pima Indian*.

Kayaer e Yildirim (2003) apresentam um sistema para diagnóstico de diabetes que foi desenvolvido usando a técnica de regressão geral de redes neurais (GRNN). GRNN aproxima qualquer função arbitrária entre a entrada e a saída, construindo uma estimativa da função, diretamente, a partir dos dados de treinamento. Logo, neste sistema, não é requerido um processo de treinamento iterativo. Este sistema obteve uma precisão de 80,21% na classificação, utilizando a base de dados *Pima Indian*.

Rajeswari e Vaithyanathan (2011) apresentam uma abordagem para auxiliar no diagnóstico de diabetes tipo 2. A abordagem construída combina a modelagem *fuzzy* com a arquitetura de RNA. Inicialmente, as informações do banco de dados são modeladas em uma abordagem *fuzzy*, sendo que esse resultado é normalizado e na sequência classificado por uma RNA *back-propagation*. Ao final o sistema proposto classifica as informações, referentes a cada pessoa, como sendo diabética ou normal. Esse sistema obteve uma taxa de eficiência de 83.3% durante a etapa de teste, em um banco de dados próprio dos pesquisadores, com 600 pessoas.

O sistema proposto, assim como os demais apresentados, utiliza uma Rede Neural para diagnóstico do diabetes. Entretanto, ao invés de utilizar a lógica *fuzzy*, emprega uma RNA *feedforward* com múltiplas camadas, que será detalhada na próxima seção, visando aumentar o índice de acertos nos diagnósticos realizados.

4. Sistema Neural para Apoio ao Diagnóstico de Diabetes

O Sistema Neural para Apoio ao Diagnóstico de Diabetes (SND) implementado, tem como pilar computacional a RNA *feedforward*, com múltiplas camadas, cujo treinamento é supervisionado, utilizando o algoritmo de treinamento *back-propagation*. A escolha deste algoritmo se deve a sua simplicidade computacional para ajuste dos pesos sinápticos e a capacidade de representar as relações não lineares e lineares. Além disso o *back-propagation* decorre da minimização do erro quadrático entre as saídas da rede e as saídas desejadas, encontrando parâmetros que conduzam o erro quadrático assintoticamente para zero. Neste artigo será detalhado o SND, conforme apresentado na Figura 1. Para desenvolver o sistema foi utilizada a linguagem de programação *Java*, juntamente com o Sistema Gerenciador de Banco de Dados *SQL Server*.

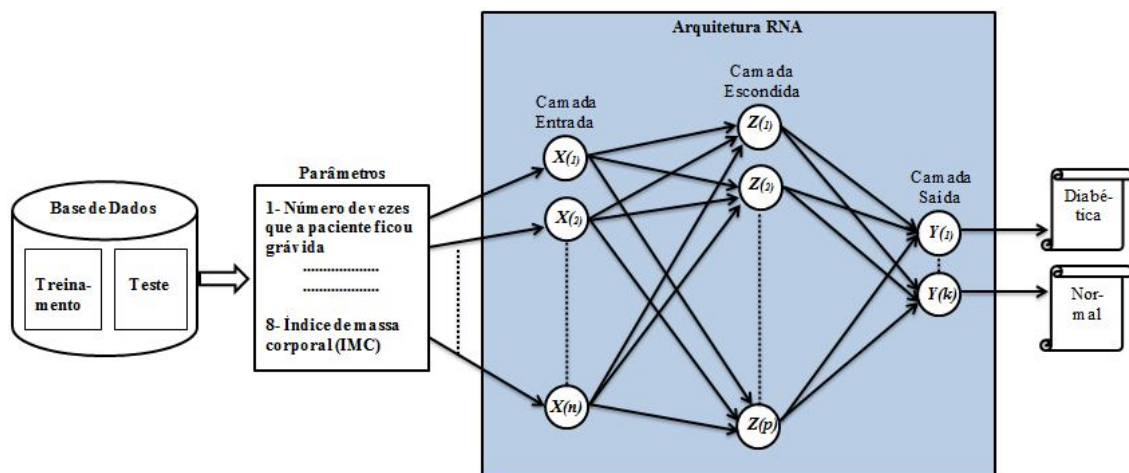


Figura 1: SND: Sistema Neural para Auxílio ao Diagnóstico de Diabetes

A base de dados utilizada tanto para treinar quanto para testar o SND foi a *Pima Indians Diabetes*(PID) (UCI, 2014), conforme apresenta, graficamente, a tabela 1, e descrito na seção 2.3.

Tabela 1. Dados de treinamento e teste da RNA

	Nb_pregnant	Plasma_glucose	Diastolic_press	Triceps_skin	serum_insulin	Diabetes	Age	Body_mass	Class
1	6	148	72	35	0	0,627	50	33,6	1
2	1	85	66	29	0	0,351	31	26,6	0
3	8	183	64	0	0	0,672	32	23,3	1
4	1	89	66	23	94	0,167	21	28,1	0

767	1	126	60	0	0	0,349	47	30,1	1
768	1	93	70	31	0	0,315	23	30,4	0

As primeiras 8 colunas (*Nb_pregnant a Body_mass*) da tabela 1 se referem às informações utilizadas para classificar cada uma das 768 pacientes. A coluna 9 informa classe a qual a paciente pertence, 1 para diabética, e 0 normal.

Para possibilitar a utilização das informações (*Nb_pregnant a Body_mass*) na RNA elas foram normalizadas. Sendo assim, cada coluna foi normalizada por meio da função: $X_{(i)} = (2 * (X_{entrada(i=1...768)} - Mínimo) / (Máximo - Mínimo)) - 1$. Logo após, foi definido o *target* bipolar para a RNA, levando em consideração as informações da coluna 9, de modo que ele tenha valor 1 para diabéticos e valor -1 para não diabéticos.

Também foram criados 2 subconjuntos oriundos de PID. Para isso foram selecionados 70% (538) dos pacientes, para o subconjunto de treinamento e os outros 30% (230 pacientes) foram utilizados para o subconjunto de teste da RNA.

De posse destas informações, a RNA foi treinada para alcançar um equilíbrio entre a capacidade de responder corretamente aos padrões de entrada, que são usados para treinamento (memorização), e a capacidade de fornecer boas respostas a entradas desconhecidas (conjunto de teste), mas semelhantes aos padrões de treinamento (generalização).

O treinamento do SND, usando o algoritmo *back-propagation* envolveu três estágios: a inserção das informações normalizadas e formatadas como entradas (*feedforward*, passos de 1 a 3) pertencentes ao conjunto de treinamento da RNA, o cálculo da retropropagação do erro (*back-propagation*, passos de 4 a 5) e o ajuste dos pesos (passo 6). A seguir é detalhado o algoritmo de treinamento em 6 passos (Fausett, 1994), conforme arquitetura apresentada, graficamente, na figura 2. A arquitetura da RNA utilizada possui 1 camada de entrada, com 8 neurônios (representados de X_1 a X_n), 1 camada oculta, composta por 10 neurônios (representados de Z_1 a Z_p) e uma de saída, com 2 neurônios (Y_1 e Y_2).

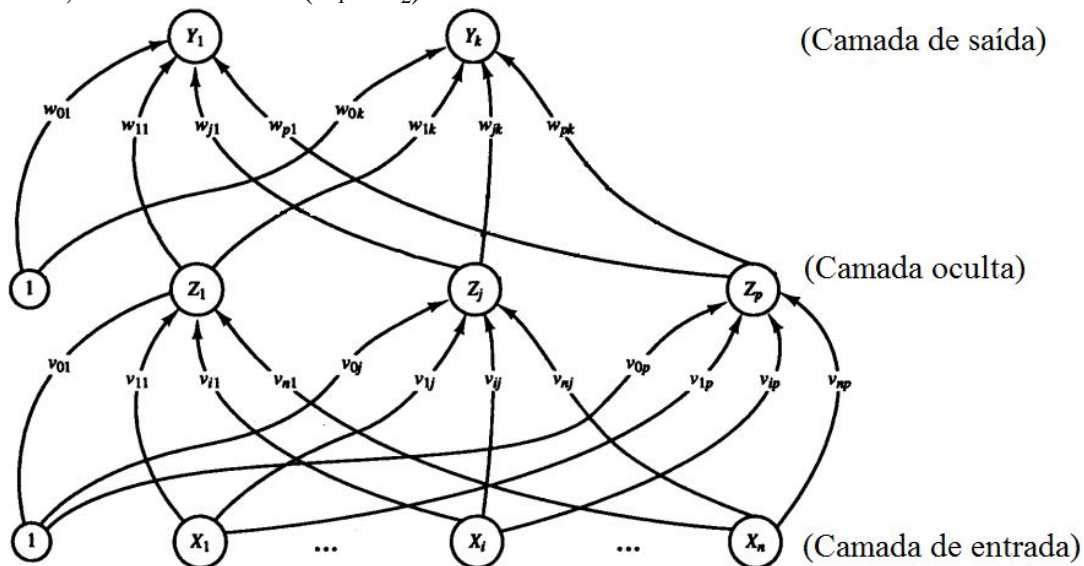


Figura 2: Arquitetura da RNA implementada

- 1) Defina a taxa de aprendizagem $\alpha = 0.03$.
- 2) Inicialize os pesos (w_0) e bias (w) com valores randômicos entre 0.5 e -0,5.
- 3) Inicie a inserção dos padrões de treinamento nas entradas da RNA. Enquanto a condição de parada for falsa, faça:
 - 3.1) Aplique o padrão de entrada $X_{(t)}$ nas camadas ocultas $Z_{(j)}$.
 - 3.2) Para cada unidade oculta ($Z_j, j=1, \dots, p$) defina $Z_{in_j} = v_{0j} + \sum_{i=1}^n x_i v_{ij}$. Aplique a função de ativação bipolar em Z_{in_j} e calcule a saída $Z_j = f(Z_{in_j})$.
 - 3.3) Para cada unidade de saída ($Y_k, k = 1, \dots, m$) defina $Z_{in_k} = w_{0k} + \sum_{j=1}^p z_j w_{jk}$.
 - 3.4) Aplique a função de ativação e calcule a saída $y_k = f(Z_{in_k})$.
- 4) Inicie a retropropagação do Erro.

4.1) Calcule a informação do erro (δ_k), para cada unidade de saída (y_k) conforme o *target* (t_k) do padrão de treinamento de entrada, logo $\delta_k = (t_k - y_k) f'(y_{in_k})$.

4.2) Calcule o termo de correção dos pesos $\Delta w_{jk} = \alpha \delta_k z_j$.

4.3) Calcule o termo de correção dos *bias* $\Delta w_{0k} = \alpha \delta_k$.

5) Para cada unidade da camada oculta ($Z_j, j=1, \dots, p$) calcule o delta das entradas $\delta_{in_j} = \sum_{k=1}^m \delta_k w_{jk}$.

5.1) Multiplique o δ_{in_j} pela derivada da função de ativação e calcule o termo da informação do erro $\delta_j = \delta_{in_j} f'(Z_{in_j})$.

5.2) Calcule os termos de correção de peso $\Delta v_{ij} = \alpha \delta_j x_i$.

5.3) Calcule os termos de correção do *bias* $\Delta v_{0j} = \alpha \delta_j$.

6) Inicie a atualização dos pesos e dos *bias*.

6.1) Para cada unidade na camada de saída ($y_k, k = 1, \dots, m$) atualize os *bias* e os pesos: $w_{jk}(\text{novos}) = w_{jk}(\text{antigos}) + \Delta w_{jk}$. Onde $j = 0, \dots, p$.

6.2) Atualize os *bias* e os pesos ($z_j, j = 1, \dots, p$), para cada unidade da camada oculta: $v_{ij}(\text{novos}) = v_{ij}(\text{antigos}) + \Delta v_{ij}$.

7) Teste a condição de parada.

Considere a função de ativação, com intervalo de [-1, 1], bipolar sigmóide como sendo $f(x) = (2/(1 + \exp(-x))) - 1$ e $f'(x) = 0.5(1 + f(x))(1 - f(x))$.

O teste do SND envolve somente a etapa de inserção das informações normalizadas e formatadas como entradas (*feedforward*, passos de 1 a 3) pertencentes ao conjunto de teste da RNA.

4.4. Validação

A validação do SND foi realizada com o subconjunto definido para a etapa de teste, sendo que estas informações, embora pertencentes ao PID, eram até então, desconhecidas pelo sistema. O total de pacientes usados para compor o teste foram 230, destes 70% são classificadas como normal e 30% diabéticas. A arquitetura da RNA utilizada possui 1 camada de entrada, com 8 neurônios e uma de saída, com 2 neurônios. O número de neurônios na camada de entrada é igual à quantidade de informações utilizadas para classificar cada paciente, ou seja, 8. Os neurônios da camada oculta foram obtidos de forma empírica. Já os da camada saída foram fixados de acordo com as possíveis classificações para cada paciente (diabética ou normal), ou seja, 2. O sistema proposto realizou 2.262 iterações para chegar aos resultados, conforme apresenta, graficamente, a figura 3.

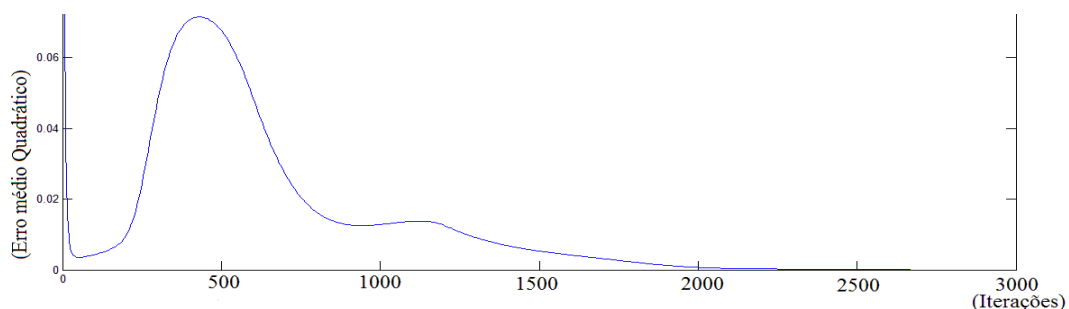


Figura 3: Gráfico do Erro Quadrático com as iterações.

O sistema proposto obteve uma taxa de acerto de 81.31%, o que resultou em 187 acertos e 43 erros (18,69%). Este resultado torna o sistema muito promissor, principalmente se comparado aos resultados obtidos pelos trabalhos apresentados em Polat (Polat et. al., 2008) e Kayaer e Yildirim (2003).

5. Considerações Finais

Este trabalho apresentou a implementação de um protótipo de sistema inteligente para apoio ao diagnóstico de diabetes, tipo 2, empregando redes neurais artificiais. Ao realizar um contraponto com trabalhos que utilizaram a mesma base de dados, PID, tais como os apresentados em Polat (Polat et. al., 2008) que obteve 82,05% de sucesso, e com Kayaer e Yildirim (2003) que obteve 80,21% de sucesso, o sistema proposto, apresenta-se como uma arquitetura promissora no diagnóstico de diabetes, pois foram alcançados 81.31% de sucesso. Alguns pesquisadores atribuem o sucesso do aprendizado da RNA (Massad, 2004) (Fausett, 1994) à técnica utilizada durante o treinamento. Neste trabalho teve-se uma atenção especial para essa etapa, a começar pela formação dos subconjuntos de treinamento e teste e, depois com a normalização das informações. Acredita-se que estes dois pontos contribuíram significativamente para o sucesso dos resultados alcançados pela RNA implementada.

Referências

- BAIER, L. J.; HANSON, R. L. **Genetic studies of the etiology of type 2 diabetes in Pima Indians**. *Diabetes*, 53, 1181–1186, 2004.
- BOTELHO, J. M. A. **Aspectos Clínicos e Epidemiológicos do Diabetes Mellitus**: Coordenadoria de Doenças e Agravos Não Transmissíveis - Superintendência de Epidemiologia/SES SUS – MG, 2003.
- FAUSETT, L. **Fundamentals of Neural Networks**: architectures, algorithms and applications, Prentice Hall International, 1994.
- HAYKIN, Simon. **Redes Neurais**: princípios e prática. 2 ed. Porto Alegre: Bookman, 2001.
- KAYAER, K.; YILDIRIM T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: **Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP)** (pp. 181–184), 2003.
- MASSAD, E; MENESES, R. X.; SILVEIRA, P. S. S; ORTEGA, N. R. S. **Métodos Quantitativos em Medicina**. Barueri, SP: Manole, 2004.
- MAZZAFERRI, E. L. **Endocrinologia**. Columbus, Ohio: Guanabara, 1988.
- PALMA NETO, L. G.; NICOLETTI, M. C. **Introdução às Redes Neurais Construtivas**. São Carlos, SP: Editora da Universidade Federal de São Carlos, 2005.
- OLIVEIRA, J. E. P.; VENCIO, S. **Diretrizes da Sociedade Brasileira de Diabetes**. São Paulo: AC Farmacêutica, 2014.
- POLAT, K.; GUNES, S.; ASLAN, A. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. **Expert Systems with Applications**, Volume 34, Issue 1, January 2008, Pages 482-487.
- RAJESWARI, K.; VAITHIYANATHAN, V. Fuzzy based modeling for diabetic diagnostic decision support using Artificial Neural Network. **IJCSNS International Journal of Computer Science and Network Security**. v.11, n.4, 2011.
- UCI. University of California, **Machine Learning and Intelligent System**, School of Information and Computer Science, 2010. Acessado em: 17/08/2014 <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>