MARCELO SCHNEIDER FIALHO

AVALIAÇÃO DA QUALIDADE DE DADOS EM PORTAIS WEB UTILIZANDO REDES NEURAIS

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de MBA em Tecnologia de Software.

São Paulo

2013

MARCELO SCHNEIDER FIALHO

AVALIAÇÃO DA QUALIDADE DE DADOS EM PORTAIS WEB UTILIZANDO REDES NEURAIS

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para a conclusão do curso de MBA em Tecnologia de Software.

Área de Concentração: Tecnologia de Software

Orientador: Prof. Dr. Jorge Rady de Almeida Junior

São Paulo

2013

FICHA CATALOGRÁFICA

Fialho, Marcelo Schneider
Avaliação da qualidade de dados em portais WEB utilizando redes neurais/ M.S. Fialho. -- São Paulo, 2013.
45 p.

Monografia (MBA em Tecnologia de Software) – Escola Politécnica da Universidade de São Paulo. Programa de Edu-cação Continuada em Engenharia.

1.Redes neurais 2.Portal 3.WEB sites I.Universidade de São Paulo. Escola Politécnica. Programa de Educação Continuada em Engenharia II.t.

DEDICATÓRIA

Dedico esta monografia a minha família pela fé e confiança demonstrada.

Aos meus amigos pelo apoio incondicional. Aos professores pelo simples fato de estarem dispostos a ensinar. Aos orientadores pela paciência demonstrada no decorrer do trabalho. Enfim a todos que de alguma forma tornaram este caminho mais fácil de ser percorrido.

AGRADECIMENTOS

Quero agradecer, em primeiro lugar, a Deus, pela força e coragem durante toda esta longa caminhada. Agradeço também todos os professores que me acompanharam durante o programa de especialização e aos meus familiares e amigos que me ajudaram e estimularam nos momentos mais difíceis.

RESUMO

A evolução da internet proporcionou o surgimento de uma grande quantidade de portais e sites por meio dos quais são oferecidos diferentes tipos de serviços e informações. Tais informações e serviços são consumidos por usuários e muitas vezes utilizados na tomada de decisões. Desta forma, para que um portal atinja seu real objetivo, é necessário que a informação ou serviço oferecido atinja as expectativas do usuário. No intuíto de atingir tais expectativas, alguns trabalhos propõem a definição de um modelo de qualidade de dados para portais web considerando a subjetividade presente na expectativa dos usuários quanto as informações oferecidas por estes portais. Este trabalho propõe a utilização de redes neurais na definição do modelo de qualidade de dados para portais web, considerando a subjetividade presente na expectativa dos usuários.

ABSTRACT

The evolution of the Internet provided biggest amount numbers of web portals wich offering differents types of services and information. Such informations and services are consumed by users and often used to make decisions. In this way, for reaches real goal, it is necessary that the information or services offered come to attend user expectations. A model of data quality for web portals is the best way to reach them. This work proposes the use of Neural Networks in the definition of data quality model for web portals.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO.

Neste capítulo serão apresentadas as motivações para este trabalho, assim como o seu objetivo e justificativas.

1.1. Motivações.

A evolução da internet proporcionou o surgimento de uma grande quantidade de portais por meio dos quais são oferecidos diferentes tipos de serviços e informações. (XIAO e DASGUPTA, 2005). Tais informações e serviços são consumidos por usuários e muitas vezes utilizados na tomada de decisões (NAM, 2009). Desta forma, para que um portal atinja seu real objetivo, é necessário que a informação oferecida atinja as expectativas e necessidades do usuário. Segundo (WAND e WANG, 1996) a má qualidade dos dados muitas vezes pode custar a credibilidade da fonte, enquanto uma boa qualidade dos dados pode ser considerada uma vantagem competitiva.

No intuíto de atingir as expectativas e necessidades dos usuários, algumas metodologias foram propostas com a finalidade de avaliar e melhorar a qualidade de dados oferecidos aos usuários (WANG, 1998). De acordo com (WAND e WANG, 1996), o conceito de qualidade de dados está relacionado ao nível de satisfação entre os dados e os requisitos de usuários. Para tanto, estudos anteriores sobre qualidade de dados definiram categorias e dimensões que permitem analisar os dados através de um conjunto de métricas (WAND e WANG, 1996) (WANG e STRONG, 1996).

Alguns trabalhos mostram que tais métricas, aplicadas junto a técnicas de inteligência artificial, possibilitaram o desenvolvimento de modelos e ferramentas que viabilizam a avaliação de uma grande quantidade de dados. É o caso de (CARO, CALERO, *et al.*, 2007) que propõe por meio de uma rede Byesiana, uma abordagem probabilística na avalição da qualidade de dados em portais. O emprego de técnicas de inteligência artificial pode também ser encontrado em (BERARDI e RUIZ, 2009) e (AL-NAMLAH e BECKER, 2009).

Diante deste cenário, o presente trabalho propõe a utilização de uma rede neural na criação de um modelo para avaliar a qualidade de dados em portais Web. Segundo (AL-NAMLAH e BECKER, 2009), redes neurais têm sido utilizadas em diferentes tipos de aplicações com diferentes propósitos como reconhecimento de padrões e processamento de sinais, e possuem a capacidade de aprender através de exemplos, podendo classificar dados

utilizando o aprendizado. O modelo proposto neste trabalho é aplicado na avaliação de qualidade dos dados de um portal web e os resultados serão apresentados.

1.2. Objetivo.

Este trabalho possui como objetivo utilizar uma rede neural na criação de um modelo para avaliar a qualidade de dados em portais Web. Este modelo é testado em um Portal Web e os resultados são apresentados.

1.3. Justificativas.

Segundo (WAND e WANG, 1996) a má qualidade dos dados muitas vezes pode influenciar na credibilidade da fonte, enquanto uma boa qualidade dos dados pode ser considerada uma vantagem competitiva. Por isso, torna-se necessária a avaliação da qualidade dos dados por meio de métricas estabelecidas de acordo com as dimensões de qualidade adequadas (WANG, 1998). Tais métricas devem ser focadas na perspectiva do usuário consumidor dos dados, que se difere das demais perspectivas – a perspectiva do produtor dos dados e a perspectiva do proprietário dos dados – pelos seguintes motivos: os usuários consumidores dos dados não possuem controle sobre a qualidade dos dados oferecidos; o objetivo do usuário consumidor é encontrar a informação que satisfaça as suas necessidades (BURGESS, GRAY e FIDDIUM, 2004) (CARO, CALERO, *et al.*, 2007).

1.4. Estrutura do trabalho.

Este trabalho está organizado em capítulos da seguinte forma. No capítulo 1 são apresentadas a motivação, o objetivo, as justificativas e a estrutura do trabalho.

O capítulo 2 apresenta conceitos sobre qualidade de dados e redes neurais. Estes conceitos visam facilitar a compreensão da metodologia adotada por este trabalho.

No capítulo 3 é apresentada a metodologia utilizada no desenvolvimento do modelo para avaliação da qualidade de dados em um portal web, assim como a aplicação prática deste modelo. Os resultados obtidos com a aplicação do modelo são apresentados no capítulo 4.

Finalmente no capítulo 5, é apresentada a conclusão sobre os resultados obtidos assim como os trabalhos futuros.

2. CONCEITOS FUNDAMENTAIS.

Neste capítulo são apresentados alguns conceitos necessários para o desenvolvimento da metodologia proposta por este trabalho. A seção 2.1 Qualidade de dados, apresenta conceitos sobre a qualidade de dados e sua aplicação, enquanto que na seção 2.2 é apresentada uma introdução sobre redes neurais.

2.1. Qualidade de dados.

Organizações têm investido cada vez mais em tecnologias que permitem efetuar a coleta, armazenamento e processamento de uma grande quantidade de dados com o propósito de aprimorar os processos de negócio, adquirir vantagens estratégicas ou tomar decisões. Porém, questões relacionadas á qualidade dos dados coletados exercem significativa influência sobre o sucesso destas tarefas. De acordo com (WAND e WANG, 1996), a má qualidade de dados pode prejudicar a credibilidade de uma organização, mas por outro lado, dados de qualidade podem representar uma vantagem competitiva. Segundo (SEBASTIAN-COLEMAN, 2013), a qualidade de dados está relacionada com o nível de satisfação das expectativas do usuário sobre os dados oferecidos. De acordo com (CAPPIELLO, FRANCALANCI e PERNICI, 2004) estes dados devem ser adequados à utilização do usuário. Desta forma, para se medir a qualidade de dados faz-se necessário entender as expectativas do usuário. Tais expectativas podem ser classificadas de acordo com diferentes perspectivas:

a) Consumidor dos dados: utilizam os dados e não possuem controle sobre a sua qualidade;
b) Proprietário dos dados: possui a responsabilidade sobre os dados. Sob esta perspectiva os usuários possuem o poder de controle sobre a qualidade dos dados;
c) Produtor dos dados: produzem os dados a fim de torná-los disponíveis para os usuários consumidores. Sob esta perspectiva os usuários possuem o controle sobre os dados produzidos, porém não podem controlar a maneira como os dados são utilizados.

Neste trabalho é considerada a perspectiva do consumidor dos dados, que difere das demais sob o aspecto de que o consumidor dos dados não possui controle sobre os dados oferecidos. Segundo (YANG, CAI, *et al.*, 2005) aplicações web possuem particularidades em relação a outros sistemas de informação, e por este motivo, (CARO, CALERO, *et al.*, 2006) sugere que a qualidade de dados deve ser focada na perspectiva do consumidor dos dados.

Quando se refere à qualidade de dados, a expectativa do usuário pode ser considerada como um conjunto de suposições relacionadas às condições dos dados. Estas expectativas podem representar a intenção de utilização dos dados ou mesmo o significado que possuem para o usuário. Com o objetivo de auxiliar a avaliação das expectativas do usuário e obter a qualidade dos dados, muitos trabalhos propõem métricas estabelecidas por meio de dimensões de qualidade. Segundo (WAND e WANG, 1996), a qualidade de dados é um conceito multidimensional e deve ser tratada como um produto com dimensões de qualidade associadas. De acordo com (SEBASTIAN-COLEMAN, 2013), um conjunto de dimensões de qualidade de dados pode ser utilizada para definir a expectativa do usuário em relação a um conjunto de dados como também avaliar a condição de um conjunto de dados já existente. Ao efetuar o estudo sobre a qualidade de dados é importante selecionar um conjunto de dimensões que reflita o contexto da aplicação e da utilização dos dados.

Embora exista uma grande quantidade de diferentes classificações para as dimensões de qualidade de dados, podem-se destacar algumas mais importantes e mais comuns na literatura como: acuracidade, integridade, consistência e temporalidade. A Tabela 1 apresenta um conjunto de categorias e suas dimensões para avaliação da qualidade de dados, sendo tal conjunto de dimensões utilizado em (CARO, CALERO, *et al.*, 2007) e obtido por meio do levantamento de características de qualidade de dados presente na literatura. Para tanto, foram considerados diferentes domínios no contexto de aplicações web como, web sites, integração de dados, comércio eletrônico, portais informativos, redes organizacionais e sistemas para tomadas de decisão.

As características obtidas foram então classificadas considerando-se dois fatores: a expectativa de qualidade de dados na perspectiva dos usuários consumidores dos dados; e as funcionalidades básicas oferecidas em portais web. Devido a sua cotextualização com aplicações web, este trabalho utiliza a dimensão representacional pesquisada em (CARO, CALERO, *et al.*, 2007) e apresentada na Tabela 1 como base para a sua proposta.

Após definida as dimensões a serem utilizadas na avaliação da qualidade dos dados é necessário estabelecer métricas para validar o resultado obtido. Por meio das métricas é possível distinguir se o resultado da avaliação da qualidade dos dados é aceitável ou não. As métricas podem ser determinadas manualmente ou de forma automática baseando-se em cálculos. Estas métricas são compostas por valores mínimos ou máximos aceitáveis. Para cada dimensão, a Tabela 2 apresenta um exemplo onde são estabelecidos valores aceitáveis para a avaliação das dimensões intrínsecas da Tabela 1.

Categoria	Dimensões
Intrínsecas: demonstra a qualidade dos dados, independente da aplicação.	• Acuracidade, • Objetividade, • Credibilidade, • Reputação.
Operacional: o sistema deve ser seguro sem comprometer seu nível de acessibilidade.	• Acessibilidade, • Segurança, • Interatividade, • Disponibilidade, • Suporte, • Facilidade de uso, • Tempo de resposta.
Contextual: conformidade entre a utilização dos dados e o seu contexto.	• Aplicabilidade, • Integridade, • Flexibilidade, • Atualização, • Confidencialidade, • Relevância, • Especialização, • Temporalidade, • Validade, • Valor agregado.
Representacional: representação dos dados de forma inteligível.	• Interpretabilidade, • Facilidade de entendimento, • Representação concisa, • Representação consistente, • Quantidade de dados, • Atratividade, • Documentação, • Organização.

Tabela 1 - Dimensões de qualidade de dados.
Fonte: autor "adaptado de" CARO, CALERO, *et al.*, 2007 p. 147

Dimensões	Valores limites
Acuracidade,	0,5 – 0,6
Objetividade,	0,2 – 0,7
Credibilidade,	0,7 – 1
Reputação.	0,8 – 1

Tabela 2 - Exemplo de métrica para as dimensões intrínsecas.
Fonte: autor

2.2. Redes Neurais Artificiais.

Nos seres humanos, a capacidade de discernimento entre características, semelhanças e padrões é efetuada de forma muito simples, porém tais tarefas são um grande desafio quando aplicadas à tecnologia. Estes desafios se devem à natureza dos problemas relacionados ao reconhecimento de padrões que, por ser considerada uma ciência não exata, dificulta a utilização de paradigmas tradicionais da computação (VALENÇA, 2010).

Por este motivo, foram desenvolvidas técnicas computacionais que se aproximam da cognição humana. Estas técnicas se baseiam em abordagens estatísticas, sintáticas, difusa ou neural e são aplicadas de acordo com o contexto e a natureza do problema. Dentre estas técnicas, a mais utilizada em tarefas que envolvem reconhecimento de padrões são as redes neurais artificiais.

As redes neurais artificiais são inspiradas nos neurônios biológicos, sendo tais neurônios representados por unidades simples de processamento que conectadas são capazes de realizar tarefas mais complexas. Estes neurônios conectados podem ser dispostos em camadas como mostra a Figura 1, onde é apresentada uma rede neural com uma camada de entrada, uma camada intermediária e uma camada de saída. As redes neurais podem ser formadas por uma ou mais camadas intermediárias, que são responsáveis pela não-linearidade e pela memória da rede (JAIN e MAO, 1996). A forma como os neurônios são dispostos e interligados definem a topologia da rede neural, as topologias necessárias no entendimento deste trabalho assim como o neurônio artificial são abordados nas seções seguintes.

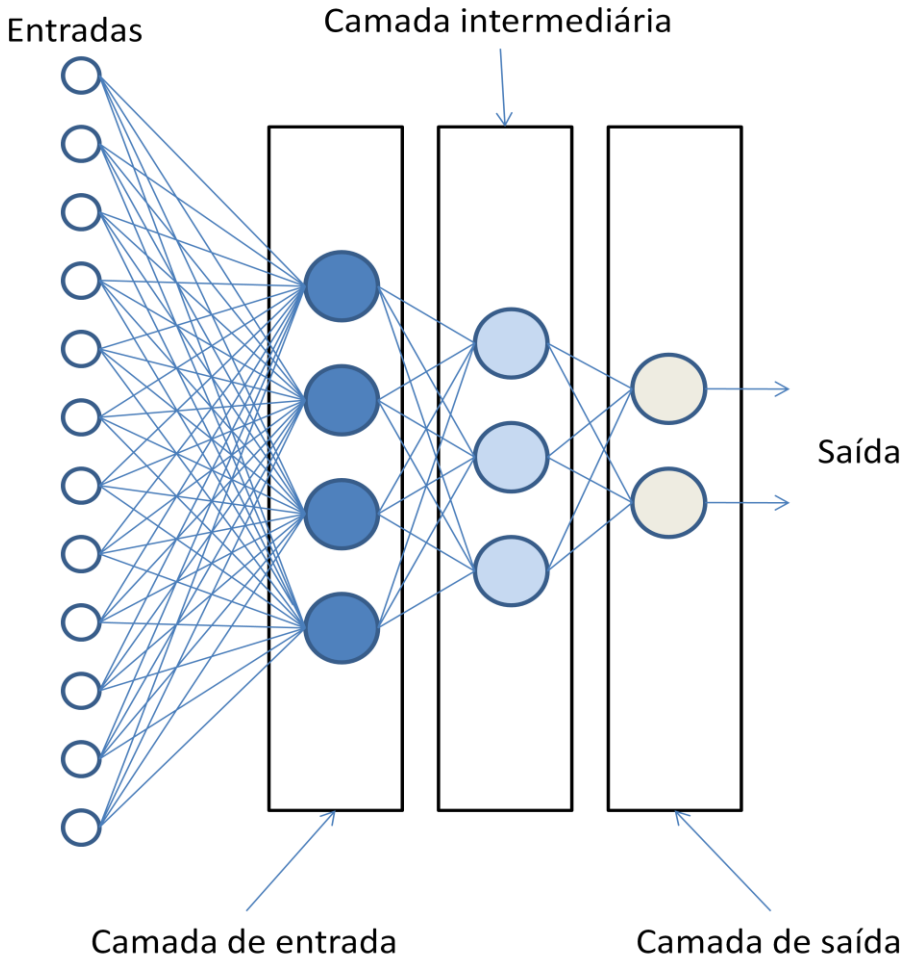

Figura 1 - Exemplo de uma rede neural.
Fonte: Autor.

2.2.1. Neurônio Artificial.

A primeira representação matemática de um neurônio biológico foi proposta por (MCCULLOCH e PITS, 1943), este modelo consiste em uma regra de propagação e função de ativação. A função de ativação do neurônio artificial, descrita pela Função ((**1**), determina um limiar para a intensidade do sinal de estimulo recebido. Desta forma, o sinal de saída de um neurônio é disparado apenas quando a intensidade de um sinal de estimulo atinge o limiar de sua função de ativação. A partir do momento em que o limiar é atingido, a intensidade do sinal de saída do neurônio permanece constante, mesmo quando o valor da intensidade do estimulo ultrapassa o limiar de sua função de ativação.

A regra de propagação é responsável pelo ajuste dos valores da intensidade do sinal de estimulo do neurônio. Esta regra é representada pela função ((**2**) onde e_i representa a intensidade do estímulo recebido pelo neurônio *i*, este estímulo é obtido através da somatória

dos produtos entre o peso w_{ij} e o valor de entrada x_j subtraído do limiar θ, para j variando de 1 até n.

$$e_i = \sum_{j=1}^{n} w_{ij}\, x_j - \theta \tag{1}$$

$$f(x_i) = \begin{cases} 1 \;\forall\; x_i \geq 0 \\ 0 \;\forall\; x_i < 0 \end{cases} \tag{2}$$

O processo de ajuste dos pesos é chamado de aprendizado. É através deste processo que um neurônio adquire a capacidade de distinguir e generalizar padrões de entrada. A Figura 1 apresenta um modelo geral de neurônio artificial com os estímulos de entrada $x_1 \ldots x_n$, e os pesos $w_1 \ldots w_n$ que são ajustados durante o processo de aprendizagem. Este modelo utiliza um limiar explícito definido por θ. Entretanto, segundo (VALENÇA, 2010), a utilização de limiares implícitos é mais adequada quando se trata de algoritmos de aprendizagem.

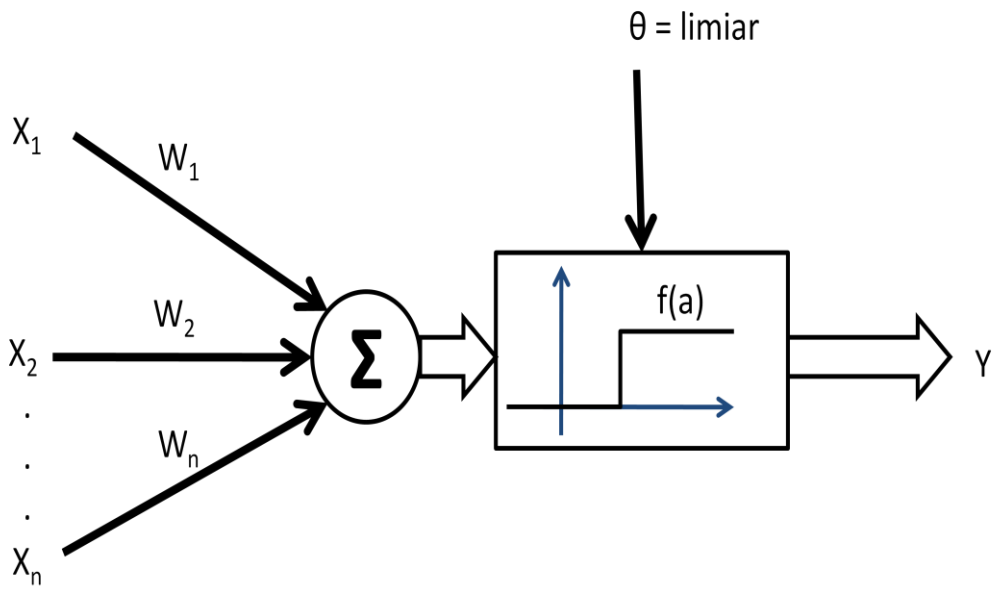

Figura 2 - Modelo geral de neurônio artificial com limiar explícito.
Fonte: Autor "adaptado de" JAIN e MAO, 1996 p. 34

O limiar implícito é caracterizado por uma entrada adicional com valor de intensidade e peso fixos. A Figura 3 apresenta um modelo de neurônio artificial com o limiar implícito x_0 e peso w_0. Figura 3

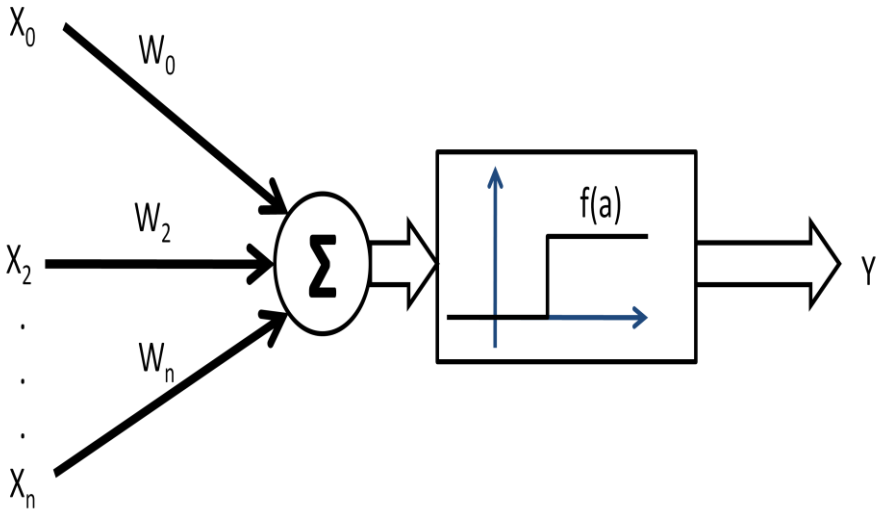

Figura 3 - Modelo de neurônio artificial com limiar implícito.
Fonte: Autor "adaptado de" JAIN e MAO, 1996 p. 34.

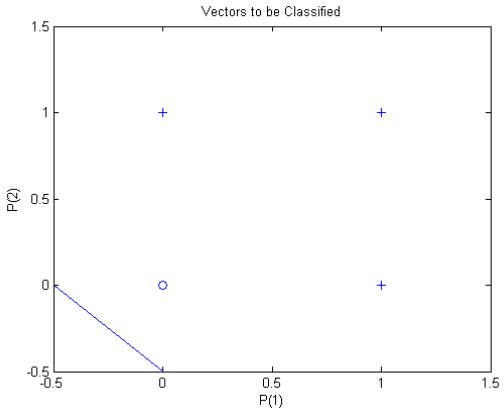

Figura 4 - Neurônio antes do treinamento.
Fonte: autor.

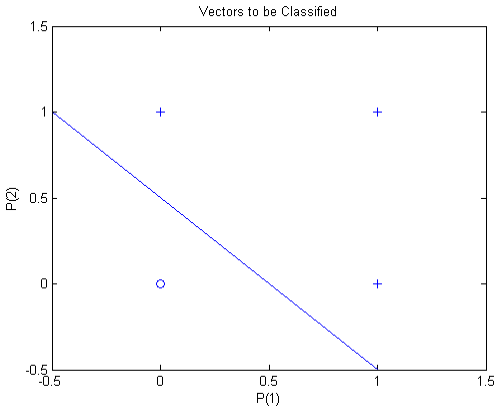

Figura 5 - Neurônio depois do treinamento.
Fonte: autor.

O ajuste do limiar implícito assim como o ajuste dos pesos auxilia a classificação de padrões apresentados como entrada para o neurônio. Esta classificação pode ser vista na Figura 4 e na Figura 5. A Figura 4 apresenta uma reta formada pelos valores assumidos no limiar antes do treinamento do neurônio. Após o treinamento os valores do limiar são ajustados e o resultado pode ser visualizado na Figura 5 onde a reta separa o ponto P(0,0) dos

demais pontos. No caso da utilização do limiar implícito, o ajuste dos pesos e do limiar é realizado através da função (3).

$$e_i = \sum_{j=0}^{n} w_{ij}\, x_i \tag{3}$$

A formulação do neurônio com a utilização do limiar implícito é a mesma utilizada pelo modelo de rede neural conhecido como perceptron. Este modelo utiliza um algoritmo de aprendizagem descrito pela função (4) onde w_{ij} representa os pesos sinápticos, α determina a taxa de aprendizagem, x_i representa o valor de entrada, d_i representa o valor de saída desejado e y_i o valor de saída obtido.

$$\Delta w_{ij} = \alpha(d_i - y_i).x_i \tag{4}$$

Por meio da Figura 4 e da Figura 5 é possível notar que o perceptron é restrito aos problemas linearmente separáveis – problemas que podem ser separados por uma reta no hiperplano. Porém, é possível resolver problemas não lineares através redes perceptron utilizando, ao menos, uma camada intermediária de neurônios como mostra a Figura 1. Esta topologia é conhecida como perceptron multi-camadas, e é apresentada na seção seguinte.

2.2.2. Perceptron Multi-camadas.

Conforme mencionado no capítulo 2.2.1 Neurônio Artificial., a rede percepetron é restrita aos problemas resolvidos através de separação linear. Desta forma, utilizando uma rede perceptron não seria possível classificar os padrões presentes na Figura 6 onde se faz necessário mais de uma reta para distinguir os pontos P(0,0) e P(1,1) dos demais. No entanto, este problema pode ser resolvido adicionando-se camadas intermediárias á rede, esta estrutura é ilustrada na Figura 7. Desta forma, a rede neural é capaz de utilizar mais de uma reta para separar diferentes padrões. Segundo (VALENÇA, 2010) uma rede perceptron multi-camadas possui as seguintes características: é formada no mínimo por três camadas, sendo uma camada de entrada de dados, uma ou mais camadas intermediárias e uma camada de saída de dados.

A camada de entrada é responsável pela apresentação dos valores de entrada á rede. O modelo apresentado pela Na Figura 7 possui três neurônios de entrada, x_1, x_2 e x_3, além do limiar x_{0+1}. Já as camadas intermediárias proporcionam a capacidade de classificar padrões de natureza não linear, estas camadas são formadas por neurônios que possuem uma função de ativação sigmoidal ou linear, na Figura 7 a camada intermediária é formada por dois neurônios e um limiar x_{0+1}.

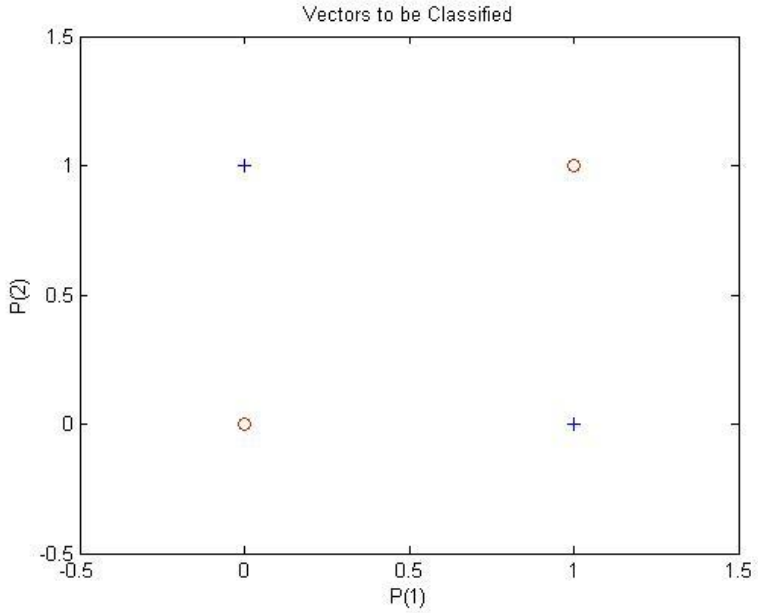

Figura 6 - Padrão inseparável através da técnica linear.
Fonte: autor.

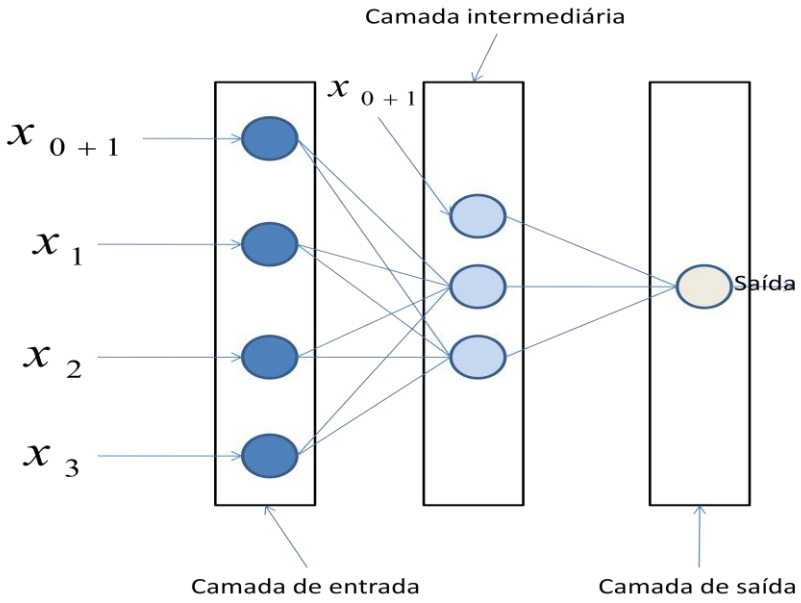

Figura 7 - Exemplo de rede perceptron multi-camadas.
Fonte: autor.

As redes perceptron multi-camadas também são caracterizadas pela forma como são treinadas. Devido à existência de camadas intermediárias utiliza-se em seu treinamento o algoritmo de backpropagation. Este algoritmo funciona de acordo com as seguintes etapas:

a) O padrão é apresentado à camada de entrada;

b) A resposta de cada neurônio da camada de entrada é propagada para a entrada de todos os neurônios da camada seguinte. Esta etapa ocorre até que se atinja a última camada da rede neural que é a camada de saída;
c) Na camada de saída, são obtidos os erros por meio da diferença entre a resposta obtida e a resposta desejada;
d) Os erros calculados na camada de saída são propagados para a camada anterior e os pesos sinápticos são ajustados considerando-se este erro. Esta etapa é realizada desde a camada de saída até a camada de entrada.

Após a apresentação do padrão á entrada da rede, é realizado o calculo da entrada líquida para a camada intermediária através da função (5), onde en_i^1 é a entrada para o neurônio i na camada 1, N_{en} representa o número de neurônios na camada de entrada, j é o índice do neurônio que emite o sinal, w_{ij} é o peso da sinapse i no neurônio j e x_j é o valor de entrada do neurônio j.

$$en_i^1 = \sum_{j=0}^{N_{en}} w_{ij} x_j \tag{5}$$

Os valores de saída dos neurônios da camada intermediária são determinados pela função de ativação sigmoide (6), onde $f^1(en_i^1)$ é a saída da camada intermediária e entrada da camada de saída.

$$f^1(en_i^1) = \frac{1}{1 + e^{-en_i^1}} \tag{6}$$

Assim como na rede perceptron, depois que o resultado da camada de saída é obtido calcula-se o erro subtraindo o valor obtido do valor esperado. Esta operação é representada pela função (7), onde err_i é o erro obtido do neurônio *i*, d_i é o valor esperado do neurônio *i* e y_i é o valor obtido para o neurônio *i*.

$$err_i = (d_i - y_i) \tag{7}$$

O erro obtido é posteriormente minimizado através do ajuste dos pesos. Este ajuste é feito com a backpropagation do erro por meio da equação (8), onde w_{ij}^m são os pesos

sinápticos a serem ajustados, α é a taxa de aprendizagem, $f^{m-1}(en_j^{m-1})$ são os sinais de entrada enviados pelos neurônios da camada anterior e δ_i^m representa o fator de sensibilidade. O fator de sensibilidade é obtido através da derivada da função de ativação de forma recursiva – da primeira para a última camada – ajustando assim os pesos sinápticos que interconectam as camadas intermediárias ás demais camadas. Esta operação é demonstrada pela função (9), onde δ_j^{m-1} representa a sensibilidade da camada anterior, $f^{m-1'}(en_j^{m-1})$ é a derivada da função de ativação.

$$w_{ij}^m(novo) = w_{ij}^m(antigo) + \alpha\delta_i^m f^{m-1}(en_j^{m-1}) \quad (8)$$

$$\delta_j^{m-1} = f^{m-1'}(en_j^{m-1}) \sum_{i=1}^{N_{net}} w_{ij}^m \delta_j^m \quad (9)$$

Durante o processo de aprendizagem, um conjunto de exemplos de entrada e seus respectivos exemplos de saída desejados são apresentados á rede. O conjunto de exemplos é repetidamente apresentado até que os valores da camada de saída satisfaçam os exemplos de saída desejados.

No capítulo seguinte é apresentado o desenvolvimento da metodologia proposta por este trabalho, onde serão aplicados os conceitos apresentados neste capítulo.

3. DESENVOLVIMENTO.

Neste capítulo é apresentada a metodologia adotada por este trabalho para medir a qualidade de dados em um portal web. A metodologia proposta consiste primeiramente na seleção dos dados. Nesta etapa são selecionadas e descritas as informações do sistema a serem avaliadas. Depois de selecionar a informação é necessário definir as dimensões de qualidade adequadas, sendo que estas dimensões são utilizadas na avaliação da informação previamente obtida. Para que a informação seja avaliada por meio das dimensões estabelecidas é necessário definir métricas que determinam o nível de qualidade da informação. Este trabalho propõe a utilização de uma rede neural para determinar o nível de qualidade obtido. Esta rede neural é treinada utilizando-se as métricas estabelecidas para cada dimensão. Após o treinamento da rede neural é realizada a validação através da aplicação do modelo aos dados armazenados no portal. Esta metodologia é representada através da Figura 8 onde cada etapa é apresentada separadamente nos capítulos seguintes.

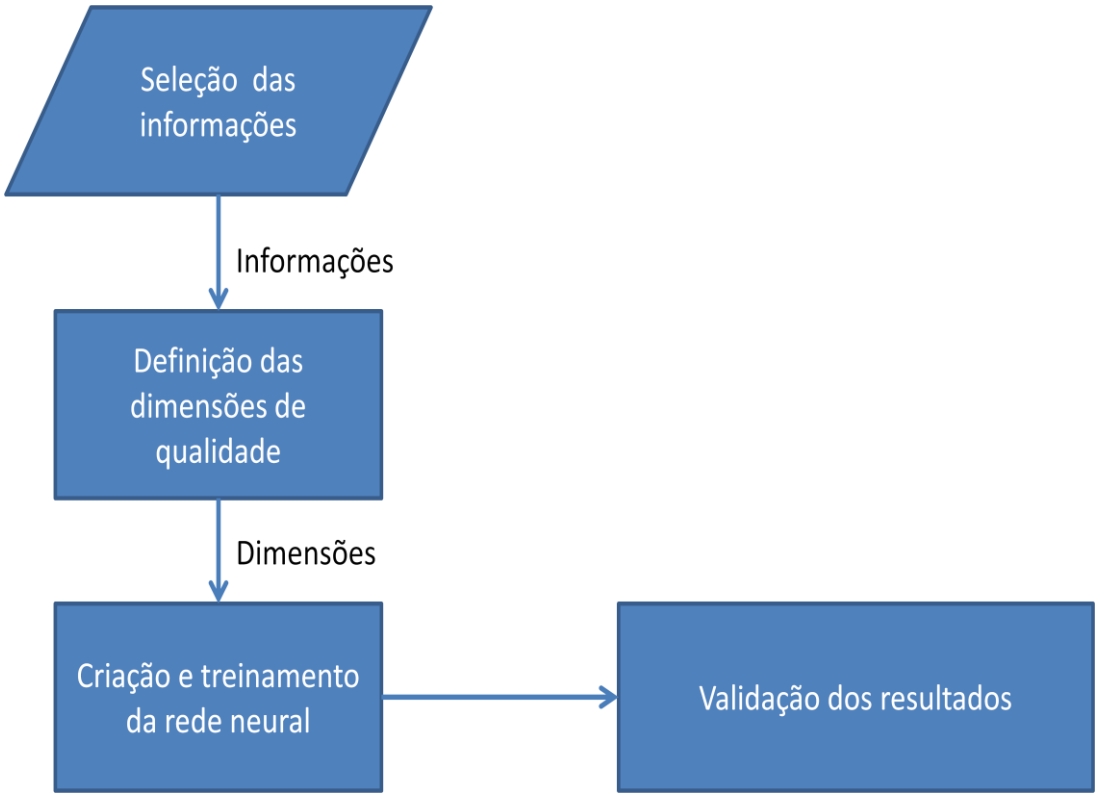

Figura 8 - Metodologia adotada para avaliação da informação.
Fonte: autor.

Na seção 3.1, é feita uma breve apresentação dos dados selecionados para a avaliação assim como o seu contexto de utilização.

3.1. Seleção das informações.

O portal analisado neste trabalho oferece ao usuário serviços como abertura de ordens de serviços, agendamento de salas de aula relatórios entre outros. Para que o usuário possa utilizar os serviços oferecidos deve cadastrar-se através de um formulário preenchendo os seguintes dados: nome; login; e-mail; senha; atividade; departamento; ramal; bloco; e sala. Os dados deste formulário são apresentados na Figura 9.

Figura 9 - Formulário de cadastro do portal.
Fonte: autor.

Os usuários cadastrados são gerenciados por meio de uma interface com opções de filtro por seção, número USP, nome e e-mail. Por meio desta interface, os usuários consumidores destes dados podem gerenciar as informações apresentadas pela Tabela 3, que também contém uma descrição sobre cada informação.

Os dados apresentados na Tabela 3 são consumidos por um grupo de usuários formado por administradores do portal e secretárias departamentais. Os administradores do portal utilizam os dados com o objetivo de controlar as permissões de acesso aos serviços oferecidos pelo portal além de extrair relatórios sobre os serviços realizados. Estes dados também auxiliam as secretárias departamentais no levantamento de alunos matriculados bem como na busca de informações de um aluno específico. As secretárias departamentais podem também controlar o acesso a áreas específicas do portal além de conceder permissões para retirada de

chaves das salas de laboratórios de acordo com o perfil descrito pelos dados dos usuários cadastrados.

Os dados cadastrais de usuários foram escolhidos para análise de qualidade devido à importância que representam para o portal e também para os usuários que os consume. O motivo da análise é também justificado pela dificuldade dos usuários consumidores dos dados com relação à administração e manipulação dos dados devido a alguns problemas como duplicidade de cadastros, interpretação incorreta e informações desatualizadas.

Dado	Descrição
Nome	Nome completo do usuário
Tipo	Papel do usuário dentro do portal. Os valores permitidos são: Usuário e Supervisor.
Login	Identificação do documento selecionado em tipo de login.
Email	Endereço de email do usuário.
Senha	Senha de acesso ao portal
Atividade	Perfil do usuário. Os valores permitidos são: aluno iniciação científica, professor, professor colaborador, funcionário, pós-doutorado e pós-graduação.
Departamento	Departamento ao qual o usuário pertence. Os valores permitidos são: Ciências Atmosféricas, Astronomia, Geofísica, Assistência técnica acadêmica, Assistência técnica administrativa e Assistência técnica financeira.
Ramal	Número do ramal telefônico do usuário.
Bloco	Bloco no qual o usuário está alocado. Os valores permitidos são: A, B, C, D, E, F, G, Área Externa, CECAS, CIentec, OBAM, ADM/BIB, Cupula e Principal.
Sala	Número da sala na qual o usuário está alocado.

Tabela 3 - Dados do formulário de cadastro do portal.
Fonte: autor.

O portal analisado neste trabalho conta atualmente com novecentos e vinte e oito cadastros, dentre os quais trezentos e trinta e dois são alunos de iniciação cientifica, duzentos e sete são pós-graduandos, quarenta e três são pós-doutorandos, oitenta e nove são professores titulares, quarenta e cinco são professores colaboradores e duzentos e doze são funcionários. Para analisar a qualidade destes dados é necessário definir as dimensões de qualidade adequadas à sua utilização e estabelecer métricas que possibilitem estabelecer o nível de

qualidade. A criação das métricas e definição das dimensões de qualidade utilizadas para a análise da qualidade dos dados cadastrais são apresentadas na seção seguinte.

3.2. Definição das dimensões de qualidade.

As dimensões de qualidade identificam aspectos sobre os dados possibilitando sua análise através de parâmetros mensuráveis. Estas dimensões devem ser adequadas à utilização dos dados considerando-se o ponto de vista dos usuários que os consomem. Neste trabalho são utilizadas as dimensões da categoria representacional estudadas por (CARO, CALERO, *et al.*, 2007). Estas dimensões foram obtidas por meio de pesquisas realizadas sobre a literatura considerando-se a expectativa de qualidade de dados na perspectiva dos usuários consumidores dos dados. Além disso, foram consideradas neste estudo, as funcionalidades básicas oferecidas em portais web. Estes fatores tornam estas dimensões adequadas para a análise da qualidade dos dados de cadastro dos usuários do portal que será analisado. A seguir são apresentadas as dimensões da categoria representacional estudadas por (CARO, CALERO, *et al.*, 2007) e utilizadas neste trabalho:

a) Facilidade de entendimento: os dados serão melhor compreendidos se forem adequadamente representados em idioma e unidades conhecidas pelos usuários.
b) Representação concisa: quando os dados são apresentados de forma objetiva e sem elementos superfulos são melhor representados;
c) Representação consistente: quando os dados possuem um formato padronizado e são compatíveis e consistentes com dados anteriores são melhor representados;
d) Quantidade de dados: a informação é melhor entendida se a quantidade ou volume de dados fornecidos for apropriado para a sua interpretação.
e) Atratividade: os dados tornam-se mais atrativos quando organizados de forma consistente através de elementos visuais.
f) Documentação: os dados são melhor compreendidos quando possuem uma documentação descritiva.

Além das dimensões descritas anteriormente, o trabalho de (CARO, CALERO, *et al.*, 2007) apresenta as dimensões organização e interpretabilidade, sendo que estas dimensões da categoria representacional foram utilizadas como uma forma de agrupar outras dimensões inter-relacionadas, formando uma estrutura de duas camadas. Neste trabalho, no entanto, não são consideradas as dimensões organização e interpretabilidade por serem análogas às dimensões atratividade e facilidade de uso, respectivamente.

Com a definição das dimensões de qualidade adequadas é possível efetuar uma análise dos dados considerando-se aspectos inerentes à sua utilização. Porém, para que seja possível obter valores mensuráveis sobre a qualidade é necessário aplicar métricas para cada dimensão.

Devido à natureza das dimensões estabelecidas, as métricas utilizadas para avaliar os dados cadastrais devem considerar a subjetividade do usuário consumidor dos dados.

Estas métricas são formadas por uma faixa de valores aceitáveis que possibilitam classificar os dados de acordo com o seu nível de qualidade. Esta faixa de valores pode ser obtida de forma automática quando se trata de métricas objetivas ou através de dados de pesquisas obtidos através de questionários respondidos pelos usuários do sistema quando se trata de métricas subjetivas.

Neste trabalho, os valores de métricas utilizados na avaliação dos dados cadastrais são fictícios e possuem o propósito de simular a metodologia aqui proposta. Para cada campo apresentado na Tabela 3 é criada uma tabela relacionando as dimensões e os valores das métricas que determinam o nível de qualidade. Os níveis de qualidade podem assumir uma faixa de valores entre zero e um e são apresentados nas seguintes tabelas: Tabela 4 apresenta métricas para o campo nome; Tabela 5 apresenta métricas para o campo tipo; Tabela 6 apresenta métricas para o campo login; Tabela 7 apresenta métricas para o campo email; Tabela 8 apresenta métricas para o campo senha; Tabela 9 apresenta métricas para o campo atividade; Tabela 10 apresenta métricas para o campo departamento; Tabela 11 apresenta métricas para o campo ramal; Tabela 12 apresenta métricas para o campo bloco; e Tabela 13 apresenta métricas para o campo sala. Nestas tabelas são apresentados, para cada dimensão, os limites mínimo e máximo para os níveis de qualidade bom, médio e ruim.

Dimensões de qualidade	Nível de qualidade para o campo nome					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,80	1	0,74	0,79	0	0,73
Representação concisa	0,90	1	0,32	0,89	0	0,31
Representação consistente	0,80	1	0,38	0,79	0	0,37
Quantidade de dados	0,90	1	0,8	0,89	0	0,79
Atratividade	0,60	1	0,45	0,59	0	0,44
Documentação	0,90	1	0,6	0,89	0	0,59

Tabela 4 - Métricas de qualidade para o campo nome.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo tipo					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,90	1	0,55	0,89	0	0,54
Representação concisa	0,90	1	0,44	0,89	0	0,43
Representação consistente	0,90	1	0,53	0,89	0	0,52
Quantidade de dados	0,70	1	0,32	0,69	0	0,31
Atratividade	0,60	1	0,47	0,59	0	0,46
Documentação	0,80	1	0,33	0,79	0	0,32

Tabela 5 - Métricas de qualidade para o campo tipo.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo login					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,90	1	0,55	0,89	0	0,54
Representação concisa	0,90	1	0,44	0,89	0	0,43
Representação consistente	0,90	1	0,53	0,89	0	0,52
Quantidade de dados	0,70	1	0,32	0,69	0	0,31
Atratividade	0,60	1	0,47	0,59	0	0,46
Documentação	0,80	1	0,33	0,79	0	0,32

Tabela 6 - Métricas de qualidade para o campo login.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo email					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,90	1	0,79	0,89	0	0,78
Representação concisa	0,70	1	0,39	0,69	0	0,38
Representação consistente	0,70	1	0,35	0,69	0	0,34
Quantidade de dados	0,90	1	0,69	0,89	0	0,68
Atratividade	0,90	1	0,45	0,89	0	0,44
Documentação	0,80	1	0,72	0,79	0	0,71

Tabela 7 - Métricas de qualidade para o campo email.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo senha					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,60	1	0,4	0,59	0	0,39
Representação concisa	0,90	1	0,5	0,89	0	0,49
Representação consistente	0,60	1	0,37	0,59	0	0,36
Quantidade de dados	0,70	1	0,33	0,69	0	0,32
Atratividade	0,70	1	0,55	0,69	0	0,54
Documentação	0,70	1	0,33	0,69	0	0,32

Tabela 8 - Métricas de qualidade para o campo senha.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo atividade					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,70	1	0,63	0,69	0	0,62
Representação concisa	0,90	1	0,33	0,89	0	0,32
Representação consistente	0,70	1	0,55	0,69	0	0,54
Quantidade de dados	0,80	1	0,44	0,79	0	0,43
Atratividade	0,90	1	0,52	0,89	0	0,51
Documentação	0,70	1	0,56	0,69	0	0,55

Tabela 9 - Métricas de qualidade para o campo atividade.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo departamento					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,70	1	0,57	0,69	0	0,56
Representação concisa	0,70	1	0,32	0,69	0	0,31
Representação consistente	0,90	1	0,58	0,89	0	0,57
Quantidade de dados	0,80	1	0,4	0,79	0	0,39
Atratividade	0,70	1	0,59	0,69	0	0,58
Documentação	0,80	1	0,35	0,79	0	0,34

Tabela 10 - Métricas de qualidade para o campo departamento.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo ramal					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,80	1	0,42	0,79	0	0,41
Representação concisa	0,90	1	0,63	0,89	0	0,62
Representação consistente	0,90	1	0,81	0,89	0	0,8
Quantidade de dados	0,80	1	0,75	0,79	0	0,74
Atratividade	0,80	1	0,68	0,79	0	0,67
Documentação	0,60	1	0,41	0,59	0	0,4

Tabela 11 - Métricas de qualidade para o campo ramal.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo bloco					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,90	1	0,73	0,89	0	0,72
Representação concisa	0,60	1	0,43	0,59	0	0,42
Representação consistente	0,80	1	0,62	0,79	0	0,61
Quantidade de dados	0,80	1	0,58	0,79	0	0,57
Atratividade	0,90	1	0,38	0,89	0	0,37
Documentação	0,80	1	0,46	0,79	0	0,45

Tabela 12 - Métricas de qualidade para o campo bloco.
Fonte: autor.

Dimensões de qualidade	Nível de qualidade para o campo sala					
	Bom		Médio		Ruim	
	Limite mínimo	Limite Máximo	Limite mínimo	Limite máximo	Limite mínimo	Limite máximo
Facilidade de entendimento	0,60	1	0,54	0,59	0	0,53
Representação concisa	0,90	1	0,39	0,89	0	0,38
Representação consistente	0,60	1	0,47	0,59	0	0,46
Quantidade de dados	0,80	1	0,43	0,79	0	0,42
Atratividade	0,90	1	0,73	0,89	0	0,72
Documentação	0,90	1	0,34	0,89	0	0,33

Tabela 13 - Métricas de qualidade para o campo sala.
Fonte: autor.

De acordo com a metodologia apresentada na Figura 8, o próximo passo consiste na criação e treinamento da rede neural utilizada para identificar o nível de qualidade dos dados avaliados. Tanto a construção da rede como o seu treinamento são apresentados na seção seguinte.

3.3. Criação e treinamento da rede neural.

A avaliação da qualidade de dados requer a utilização de métricas através das quais seja possível determinar o seu nível de qualidade. Na seção 3.2, foram definidas as métricas para avaliação de cada um dos dados cadastrais apresentados pela Tabela 3. Nesta seção é apresentada a construção de uma rede neural capaz de aplicar aos dados cadastrais as métricas definidas no capítulo anterior. Para tanto, é necessário determinar a topologia de rede neural adequada e posteriormente, efetuar o seu treinamento para que seja capaz de distinguir os níveis de qualidade dos dados. Tanto a definição da topologia quanto o treinamento da rede neural são abordados nas seções seguintes.

3.3.1. Definição da topologia da rede neural.

A topologia de uma rede neural determina a quantidade de camadas intermediárias utilizadas bem como a quantidade de neurônios utilizados em cada camada. Para a metodologia adotada neste trabalho, a utilização da rede neural perceptron multi-camadas é mais adequada devido à sua capacidade de classificação não linear. A topologia adotada para a rede perceptron multi-camadas é formada por uma camada de entrada com doze neurônios, sendo uma entrada para cada campo do formulário, um neurônio para o tipo de dimensão e um neurônio para o limiar bias. Esta topologia possui ainda duas camadas intermediárias – a primeira formada por cinco neurônios e a segunda formada por quatro neurônios – e uma camada de saída formada por um neurônio. Esta topologia é ilustrada pela Figura 10.

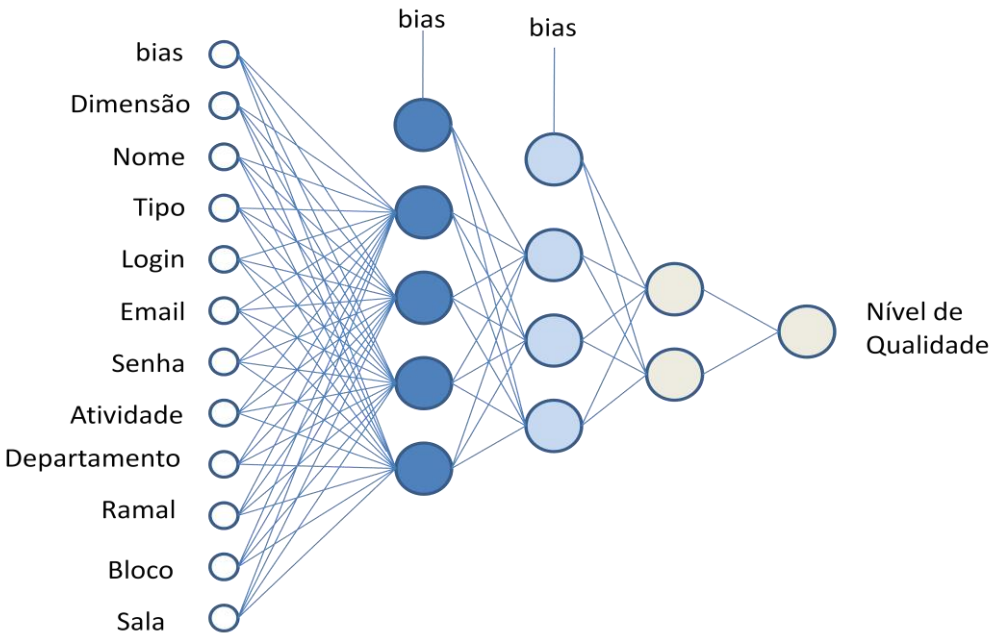

Figura 10 - Topologia da rede neural utilizada.
Fonte: autor.

Ainda na Figura 10 é possível notar na camada de entrada e nas camadas intermediárias a presença de um neurônio de entrada denominado bias. Este neurônio funciona como um limiar e auxilia nos ajustes dos pesos sinápticos e na classificação de diferentes tipos de padrões. A quantidade de camadas intermediárias é determinada de forma empírica através da realização de testes. Vale a pena destacar que foi criada apenas uma rede neural para todos as dimensões de qualidade da Tabela 13.

Após a determinação da topologia da rede neural faz-se necessário o seu treinamento por meio de valores de exemplo e seus respectivos resultados desejados. Depois de treinada a rede neural é capaz de apresentar e até generalizar valores de respostas desejados de acordo com os valores de entradas oferecidos. O treinamento da rede neural é apresentado na seção seguinte.

3.3.2. Treinamento da rede neural.

O processo de treinamento de uma rede neural consiste no ajuste dos valores dos pesos que interligam seus neurônios. Assim, valores de exemplo são apresentados na entrada da rede e a saída obtida é comparada com a saída desejada. Quando o valor obtido se difere do valor desejado os pesos são ajustados e os valores são novamente apresentados. O ciclo que compreende a apresentação dos valores de entrada e posteriormente o ajuste dos pesos é

chamado de época, sendo este processo continuamente repetido até que os pesos sejam ajustados de forma adequada a oferecer a saída desejada.

O ajuste dos pesos que interligam os neurônios é realizado durante o processo de treinamento por meio de um algoritmo. O algoritmo mais utilizado no treinamento de redes neurais perceptron multi-camadas é chamado de backpropagation. Este algoritmo identifica o erro obtido na camada de saída e propaga este erro para as camadas anteriores ajustando seus pesos. Este algoritmo é utilizado para o treinamento da rede neural definida na seção 3.3.1,seu funcionamento é descrito na seção 2.2.2.

Para realizar o treinamento da rede neural definida na seção 3.3.1, foram estabelecidos valores fictícios para o nível de qualidade que seja deseja obter como meta na avaliação dos dados do formulário. Estes valores representam o nível mínimo aceitável para que o formulário avaliado seja considerado bom, médio ou ruim de acordo com uma determinada dimensão. Desta forma, de acordo com a Tabela 14, para que o formulário possua o nível de facilidade de entendimento bom, é necessário que os dados, nome, tipo, login, senha, bloco e sala sejam avaliados com o nível de qualidade bom, enquanto os dados, e-mail, atividade, departamento e ramal devem ser avaliados no mínimo com o nível de qualidade médio. Os valores para o nível de qualidade desejado são apresentados na Tabela 14 e definem os padrões de qualidade para o formulário de cadastro.

Dimensões de qualidade	Nome	Tipo	Login	Email	Senha	Atividade	Departamento	Ramal	Bloco	Sala	Qualidade do formulário
Facilidade de entendimento	Bom	Bom	Bom	Médio	Bom	Médio	Médio	Médio	Bom	Bom	**Bom**
Representação concisa	Bom	Bom	Bom	Médio	Bom	Médio	Médio	Médio	Médio	Médio	
Representação consistente	Bom	Médio	Médio	Bom	Médio	Bom	Bom	Bom	Bom	Bom	
Quantidade de dados	Médio	Médio	Médio	Médio	Médio	Bom	Bom	Bom	Médio	Bom	
Atratividade	Bom	Bom	Bom	Médio	Bom	Bom	Bom	Bom	Médio	Médio	
Documentação	Médio	Bom	Bom	Médio	Médio	Bom	Bom	Médio	Médio	Médio	
Facilidade de entendimento	Médio	Médio	Médio	Médio	Médio	Médio	Ruim	Médio	Médio	Médio	**Médio**
Representação concisa	Médio	Médio	Médio	Ruim	Médio	Médio	Ruim	Ruim	Médio	Médio	
Representação consistente	Médio	Médio	Ruim	Ruim	Médio	Médio	Ruim	Médio	Médio	Médio	
Quantidade de dados	Médio	Médio	Ruim	Ruim	Ruim	Médio	Médio	Médio	Médio	Ruim	
Atratividade	Ruim	Médio	Médio	Ruim	Ruim	Médio	Médio	Médio	Médio	Médio	
Documentação	Médio	Médio	Ruim	Ruim	Médio	Médio	Médio	Médio	Ruim	Médio	
Facilidade de entendimento	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	**Ruim**
Representação concisa	Ruim	Médio	Ruim	Ruim	Ruim	Ruim	Médio	Ruim	Ruim	Ruim	
Representação consistente	Ruim	Ruim	Médio	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	
Quantidade de dados	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	
Atratividade	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Médio	Ruim	Ruim	Ruim	
Documentação	Ruim	Ruim	Médio	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	

Tabela 14 – Níveis de qualidade desejados para o formulário.
Fonte: autor.

Para se utilizar os dados da Tabela 14 no treinamento da rede neural torna-se necessário parametrizar os seus valores. Esta parametrização tem por objetivo facilitar o treinamento da rede neural e é feita com a definição de um valor numérico correspondente para cada dimensão de qualidade e com a definição de um valor numérico correspondente para cada valor de nível de qualidade – Bom, Médio, Ruim. A definição dos valores numéricos para as dimensões de qualidade são apresentadas na Tabela 15 enquanto a definição dos valores numéricos para os níveis de qualidade são apresentados na Tabela 16.

Dimensões de qualidade	Valores representativos
Facilidade de entendimento	1
Representação concisa	2
Representação consistente	3
Quantidade de dados	4
Atratividade	5
Documentação	6

Tabela 15 - Valores numéricos para dimensões de qualidade.
Fonte: autor.

Nível de qualidade	Valor representativo
Bom	3
Médio	2
Ruim	1

Tabela 16 - Valores numéricos para níveis de qualidade.
Fonte: autor.

Com a parametrização dos valores de nível de qualidade e das dimensões de qualidade é possível obter um conjunto de valores de entrada para a rede neural apresentada na Figura 10. Este conjunto de valores de entrada é obtido substituindo-se os valores da Tabela 14 pelos correspondentes valores parametrizados da Tabela 15 e da Tabela 16. Esta substituição é apresentada na Tabela 17. É possível notar por meio desta tabela que, tanto as entradas com os valores mínimos como os possíveis valores máximos fazem parte dos valores de entrada para o treinamento da rede neural.

Nome	Tipo	Login	Email	Senha	Atividade	Departamento	Ramal	Bloco	Sala	Dimensão	Qualidade
3	3	3	3	3	3	3	3	3	3	1	
3	3	3	3	3	3	3	3	3	3	2	
3	3	3	3	3	3	3	3	3	3	3	
3	3	3	3	3	3	3	3	3	3	4	
3	3	3	3	3	3	3	3	3	3	5	
3	3	3	3	3	3	3	3	3	3	6	**Bom**
3	3	3	2	3	3	3	3	3	3	1	
3	3	3	2	3	3	3	3	3	3	2	
3	2	2	3	2	3	3	3	3	3	3	
2	2	2	2	2	3	3	3	2	3	4	
3	3	3	2	3	3	3	3	2	2	5	
2	3	3	2	2	3	3	2	2	2	6	
2	2	2	2	2	2	2	2	2	2	1	
2	2	2	2	2	2	2	2	2	2	2	
2	2	2	2	2	2	2	2	2	2	3	
2	2	2	2	2	2	2	2	2	2	4	
2	2	2	2	2	2	2	2	2	2	5	
2	2	2	2	2	2	2	2	2	2	6	
2	2	2	2	2	2	1	2	2	2	1	**Médio**
2	2	2	1	2	2	1	1	2	2	2	
2	2	1	1	2	2	1	2	2	2	3	
2	2	1	1	1	2	2	2	2	1	4	
1	2	2	1	1	2	2	2	2	2	5	
2	2	1	1	2	2	2	2	1	2	6	
1	1	1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	1	2	
1	1	1	1	1	1	1	1	1	1	3	
1	1	1	1	1	1	1	1	1	1	4	
1	1	1	1	1	1	1	1	1	1	5	
1	1	1	1	1	1	1	1	1	1	6	**Ruim**
1	2	1	1	1	1	2	1	1	1	2	
1	1	2	1	1	1	1	1	1	1	3	
1	1	1	1	1	1	2	1	1	1	5	
1	1	2	1	1	1	1	1	1	1	6	

Tabela 17 - Valores de treinamento para a rede neural.
Fonte: autor.

O processo de treinamento da rede neural é realizado apresentando-se linha por linha da Tabela 17 à sua entrada. Este processo é repetido para cada linha da tabela até que se obtenha o valor de saída desejado para cada conjunto de entradas. O treinamento da rede é simulado utilizando-se o software Matlab por meio do seguinte código:

O processo de treinamento da rede neural por meio dos valores da Tabela 17 é apresentado pelo gráfico da Figura 11. Neste gráfico é possível verificar que foram necessárias duzentos e setenta e uma épocas para que os resultados desejados fossem obtidos com o menor erro possível. Tanto a simulação da rede neural como o seu treinamento foram realizados por meio do software Matlab.

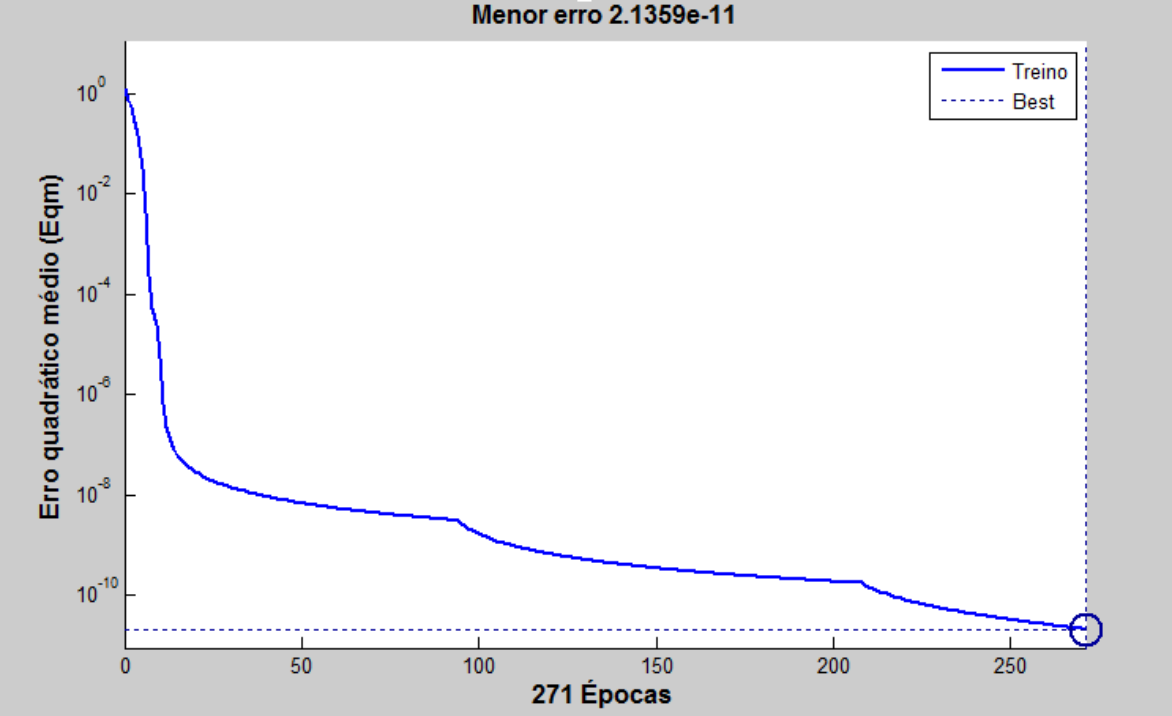

Figura 11- Treinamento da rede neural.
Fonte: autor.

Com a rede neural treinada é possível utiliza-la para determinar o nível de qualidade dos dados do formulário apresentado na seção 3.1. Na seção seguinte, os dados utilizados no treinamento são validados por meio da utilização da rede neural. Após a validação a rede neural é utilizada para avaliar os dados da Tabela 3.

3.4. Validação dos resultados obtidos através da rede neural.

Antes da utilização da rede neural para a avaliação da qualidade dos dados do formulário, é realizada a validação de seu funcionamento. Para verificar se a rede está

devidamente treinada, primeiramente é necessário utilizar como entrada os conjuntos de valores utilizados em seu treinamento apresentados na Tabela 17. Ao utilizar estes valores, deve-se obter como saída os valores de nível de qualidade pré-estabelecidos durante o processo de treinamento para cada conjunto. A Tabela 18 mostra os resultados obtidos por meio desta verificação para alguns dos conjuntos de treinamento da Tabela 17. Estes resultados são semelhantes aos resultados desejados, isto significa que a rede neural está treinada para avaliar o formulário de acordo com os níveis de qualidade estabelecidos pela Tabela 14.

Dimensões de qualidade	**Nome**	**Tipo**	**Login**	**Email**	**Senha**	**Atividade**	**Departamento**	**Ramal**	**Bloco**	**Sala**	**Obtido**	**Esperado**
Facilidade de entendimento	Bom	Bom	Bom	Médio	Bom	Médio	Médio	Médio	Bom	Bom	Bom	Bom
Facilidade de entendimento	Médio	Médio	Médio	Médio	Médio	Médio	Ruim	Médio	Médio	Médio	Médio	Médio
Facilidade de entendimento	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim	Ruim

Tabela 18 - Validação da rede neural.
Fonte: autor.

No capítulo seguinte é realizada a avaliação do nível de qualidade do formulário de cadastro, além da análise dos resultados obtidos.

4. ANÁLISE DOS RESULTADOS.

Com a rede neural treinada é possível avaliar o nível de qualidade do formulário para cada dimensão da classe representacional. Para tanto, primeiramente é necessário obter a avaliação de cada dado do formulário de acordo com as dimensões de qualidade estabelecidas. Esta avaliação deve ser obtida por meio de um questionário respondido pelos usuários consumidores dos dados. No entanto, neste trabalho, estes dados são simulados com a finalidade de verificar o funcionamento da metodologia proposta. Estes dados são apresentados pela Tabela 19 e representam as notas atribuídas pelo usuário para dado do formulário de acordo com a dimensão de qualidade.

Dimensões	Nome	Tipo	Login	Email	Senha	Atividade	Departamento	Ramal	Bloco	Sala
Facilidade de entendimento	0,64	0,21	0,13	0,16	0,87	0,91	0,97	0,67	0,03	0,2
Representação concisa	0,42	0,38	0,55	0,99	0,74	0,11	0,22	0,4	0,82	0,19
Representação consistente	0,75	0,16	0,06	0,89	0,31	0,72	0,58	0,74	0,46	0,05
Quantidade de dados	0,59	0,04	0,64	0,74	0,04	0,97	0,1	0,28	0,74	0,43
Atratividade	0,09	0,03	0,26	0,03	0,74	0,25	0,81	0,73	0,86	0,05
Documentação	0,41	0,47	0,83	0,18	0,67	0,2	0,7	0,81	0,74	0,3

Tabela 19 - Simulação de pesquisa com usuário.
Fonte: autor.

Para que os dados simulados possam ser utilizados para avaliar o formulário por meio da rede neural, é necessário que sejam classificados de acordo com as métricas estabelecidas na seção 3.2. O resultado desta classificação é apresentado na Tabela 20. Após a classificação, os dados simulados já podem ser utilizados na rede neural. A Tabela 21 apresenta os níveis de qualidade obtidos para a avaliação do formulário por meio da rede neural. Estes resultados são analisados no capítulo seguinte.

Dimensões	Nome	Tipo	Login	Email	Senha	Atividade	Departamento	Ramal	Bloco	Sala
Facilidade de entendimento	Ruim	Ruim	Ruim	Ruim	Bom	Bom	Bom	Médio	Ruim	Ruim
Representação concisa	Médio	Ruim	Médio	Bom	Médio	Ruim	Ruim	Ruim	Bom	Ruim
Representação consistente	Médio	Ruim	Ruim	Bom	Ruim	Bom	Médio	Ruim	Ruim	Ruim
Quantidade de dados	Ruim	Ruim	Médio	Médio	Ruim	Bom	Ruim	Ruim	Médio	Médio
Atratividade	Ruim	Ruim	Ruim	Ruim	Bom	Ruim	Bom	Médio	Médio	Ruim
Documentação	Ruim	Médio	Bom	Ruim	Médio	Ruim	Médio	Bom	Médio	Ruim

Tabela 20 - Classificação de acordo com as métricas.
Fonte: autor.

Dimensões	Nível de qualidade obtido
Facilidade de entendimento	Bom
Representação concisa	Ruim
Representação consistente	Ruim
Quantidade de dados	Ruim
Atratividade	Bom
Documentação	Bom

Tabela 21 - Resultados obtidos por meio da rede neural.
Fonte: autor.

5. CONSIDERAÇÕES FINAIS.

Neste trabalho foi proposta a utilização de uma rede neural na identificação do nível de qualidade de um portal web. Devido à particularidade apresentada pelos portais web em relação aos outros sistemas computacionais, foram utilizadas dimensões de qualidade focadas no usuário consumidor dos dados. Essas dimensões permitem avaliar questões como a facilidade de uso, atratividade, representação e consistência dos dados.

Com o intuito de colocar em prática a metodologia proposta, neste trabalho foram selecionados os dados do formulário de cadastro de um portal web. Para efetuar a análise de qualidade destes dados foram definidas métricas de qualidade que permitiram classificar os dados de acordo com os níveis de qualidade bom, médio e ruim. Estas métricas foram baseadas no com a definição de um limiar para os níveis de qualidade bom médio e ruim. Por meio desta classificação foi realizado o treinamento da rede neural utilizando o algoritmo backpropagation que demonstrou resultados satisfatórios na identificação do nível de qualidade do formulário analisado para a classe de dimensão representacional, podendo ser aplicado para a análise de dados reais.

Esta metodologia permite acompanhar o nível de qualidade dos dados onde é possível definir um limiar para os níveis de qualidade bom, médio e ruim.

Embora os resultados obtidos tenham sidos satisfatórios para as análises realizadas, em casos particulares a rede não é capaz de generalizar valores para os níveis de qualidade de forma adequada. Como trabalhos futuros, propõem-se a utilização de lógica fuzzy em conjunto com a rede neural, de forma a obter resultados mais apurados, além da utilização das outras classes de dimensões.

BIBLIOGRAFIA

AL-NAMLAH, A.; BECKER, S. A. Employing Neural Networks to Assess Data Quality. **Issues & Trends of Information Technology Management in Contemporary Organizations**, Hershey, p. 28-32, 2009.

BERARDI, R. C. G.; RUIZ, D. D. A. **Fuzzy-Provenance Architecture for Effort Metric Data Quality Assessment**. Rio Grande do Sul: [s.n.]. 2009.

BURGESS, M. S. E.; GRAY, W. A.; FIDDIUM, N. J. **Quality Measures And The Information Consumer**. Ninth International Conference on Information Quality. Wales: [s.n.]. 2004. p. 373-388.

CAPPIELLO, C.; FRANCALANCI, C.; PERNICI, B. **Data quality assessment from the user's perspective**. Proceedings of the 2004 international workshop on Information quality in information systems. Nova York: ACM. 2004. p. 68-73.

CARO, A. et al. **Defining a Data Quality Model for Web Portals**. ICWE '06 Proceedings of the 6th international conference on Web engineering. New York: ACM. 2006. p. 115-116.

CARO, A. et al. **A Probabilistic Approach to Web Portal's Data Quality Evaluation**. Quality of Information and Communications Technology, 2007. QUATIC 2007. 6th International Conference on the. Lisboa: [s.n.]. 2007. p. 143 - 153.

EPPLER, M. J. **Managing Information Quality:** Increasing the Value of Information in Knowledge-intensive Products and Processes. 2ª Edição. ed. [S.l.]: Springer, 2006.

HUANG, K.-T.; WANG, R. Y.; LEE, Y. W. **Quality Information and Knowledge Management**. Rio de Janeiro: Prentice-Hall do Brasil Ltda., 1999.

JAIN, A. K.; MAO, J. Artificial Neural Networks: A Tutorial. **Computer**, Wilmington, v. 29, n. 3, p. 31-44, Março 1996.

MCCULLOCH, W. S.; PITS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 115-133, 1943.

NAM, J. **Web portal quality**. Service Operations, Logistics and Informatics, 2009. Chicago: [s.n.]. 2009. p. 163 - 168.

SEBASTIAN-COLEMAN, L. **Measuring Data Quality for Ongoing Improvement A Data Quality Assesment Framework**. Waltham: Elsevier, 2013.

VALENÇA, M. J. S. **Fundamentos das Redes Neurais:** exemplos em Java. 2ª Edição. ed. Recife: Livro Rápido, 2010.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological. **Communications of the ACM**, New York, v. 39, p. 86-95, novembro 1996.

WANG, R. Y. A product perspective on total data quality management. **Communications of the ACM**, New York, v. 41, p. 58-65, Fevereiro 1998.

WANG, R. Y.; LEE, Y. W.; STRONG, L. L. P. A. D. M. Manage Your Information as a Product. **Sloan Management Review**, p. 95-105, Julho 1998.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, Armonk, Março 1996. 5-33.

XIAO, L.; DASGUPTA, S. User Satisfaction with Web Portals: An Empirical Study. In: GAO, Y. **Web Systems Design and Online Consumer Behavior**. Hershey: Idea Group Publishing, 2005. Cap. 11, p. 193-205.

YANG, Z. et al. Development and validation of an instrument to measure user perceived service quality of information presenting web portals. **Information and Management**, Kowloon, Hong Kong, v. 42, n. 4, p. 575-589, Maio 2005.

APÊNDICE 1 – CÓDIGO PARA A SIMULAÇÃO DA REDE NEURAL NO MATLAB.

A rede neural foi criada e treinada utilizando o software Matlab por meio do seguinte código:

```
>> rede = feedforwardnet([4 3 2]);
>> In = Base de treinamento;
>> out = Saídas desejadas para a base de treinamento;
>> rede.divideFcn=' ';
>> rede = train(rede,In,out);
```