colloquium exactarum

EXTRAÇÃO DE CONHECIMENTO EM BASE DE DADOS DAS MEDIDAS DOS PARÂMETROS FÍSICOS DA REDE DO ASSINANTE

EXTRACTION OF KNOWLEDGE DATA BASED ON PHYSICAL PARAMETER MEASURES OF THE SUBSCRIBER'S NETWORK

Edson Brito Júnior

Universidade Federal do Pará - UFPA, Programa de Pós Graduação em Engenharia Elétrica, Campus Universitário do Guamá, PA
E-mail: edbjr@ufpa.br

RESUMO – Com a popularização da rede mundial de computadores, a Internet, informações no formato digital tornam-se mais acessíveis. Documentos, arquivos de imagem, áudio e vídeo trafegam em todo instante na rede. Novos aplicativos surgem na rede, entre eles destacam-se as conferências, descarregamento de vídeo, jogos etc. Dentro deste contexto surge a necessidade de se alcançar taxas de bits maiores do que as fornecidas pela conexão tradicional discada, para que este serviço seja oferecido com excelente qualidade ao assinante, é de fundamental interesse para as operadoras conhecer o estado do enlace telefônico deste assinante. Com intuito de qualificar o enlace do assinante, propõe-se neste trabalho experimentos com Mineração de Dados para extrair conhecimentos das informações armazenadas em uma base de dados referentes às medidas dos parâmetros físicos da rede dos assinantes.
Palavras-chave: base de dados; extração de conhecimento; rede do assinante.

ABSTRACT – The computers network popularization, the Internet, some information in digital format becomes more accessible. Documents, image files, sound and video pass through all time in the network. New applications appear, it's distinguished among them the video conferences, and video uploads, games. Inside this context appears the necessity to get a larger transference of bits offered by the traditional dial-up connection, to offer this service in a high quality for the subscribers, its essential for the telephone companies know the status of subscribers' telephone loop. With the intention to qualify the subscribers' enlace, the suggestion is an implantation of a methodology through knowledge extraction, using the stored information in database of measures subscribers' loop.
Keywords: data base; extraction knowledge; subscriber's network.

Recebido em: 19/04/2017
Revisado em: 05/09/2017
Aprovado em: 20/10/2017

1. INTRODUÇÃO

Com o constante surgimento de novos aplicativos na Internet, e com a grande evolução do processamento de dados, torna-se cada vez mais difícil manter o gerenciamento e qualidade dos serviços e da infraestrutura dessa rede. E dentre as soluções encontradas para acesso à Internet em alta velocidade, destaca-se a tecnologia DSL (*Digital Subscriber Line*) que utiliza a infraestrutura telefônica existente.

Uma maneira de avaliar o estado atual do enlace é conhecer os valores de medições dos parâmetros físicos conhecidos como: Função de Transferência, Impedância de Entrada e Parâmetro de Espalhamento S_{11}, com isso as operadoras podem chegar a um resultado mais preciso sobre a atual condição do enlace local do assinante. Essa avaliação é conhecida como Qualificação do enlace (*Loop Qualification*). Diante disso, surgem inúmeras possibilidades de utilizar as informações armazenadas na base de dados para extrair conhecimento e tomada de decisões quanto ao melhoramento e ao aumento da taxa de bits entregue ao assinante.

O principal objetivo deste artigo é apresentar experimentos com o uso da técnica de inteligência computacional usando uma base de dados de medições dos parâmetros físicos do enlace local do assinante, ressaltando a sua importância na utilização desses modelos para se classificar um enlace com presença ou não de extensões (classe binária).

Este artigo está organizado da seguinte maneira: na Seção 2, são revisados os conceitos importantes sobre o enlace local do assinante e da mineração de dados. Na Seção 3, é descrito a metodologia para os experimentos de classificação binária. Na seção 4, são descritos os experimentos realizados. Na seção 5, é visto a análise dos resultados para o teste dos classificadores e Finalmente, na seção 6 são apresentadas as considerações finais.

2. FUNDAMENTOS

2.1. Enlace Local do Assinante

Segundo (MUNCINELLI, 2001) o enlace telefônico de par trançado é chamado de enlace local do assinante. Sendo assim, pelo fato de possuir várias topologias de comprimentos variáveis e muitas vezes maiores que 3 km são constituídos de uma ou mais seções e extensões. Uma extensão (*bridge tap*) é qualquer comprimento de cabo que não esteja dentro do caminho central telefônica/modem do assinante. Por exemplo, um par utilizado anteriormente para outro aparelho telefônico, que continua conectado em uma posição intermediária. A existência de uma extensão não é fator determinante para o impedimento de um serviço DSL, mas sim o seu comprimento.

2.2. Mineração de Dados

Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) Mineração de Dados é um conjunto de técnicas computacionais para a extração de informações desconhecidas e potencialmente úteis em grandes volumes de dados por meio de um resumo compacto dos mesmos. O termo “mineração de dados” é uma das etapas de um processo maior denominado descoberta de conhecimento em banco de dados *(KDD - Knowledge Discovery in Databases (KDD)*.

Os métodos de Mineração de Dados requerem diferentes necessidades de pré-processamento (MORIK, 2001). Tais necessidades variam em função do aspecto extensional da base de dados em que o método será utilizado. Em decorrência da grande diversidade de métodos de pré-processamento de dados, são muitas as alternativas possíveis de combinações entre métodos. A escolha dentre estas alternativas pode influenciar na qualidade do resultado do processo de KDD (MORIK, 2001).

Segundo (GOLDSCHMIDT; PASSOS, 2017) a expressão “método baseado em instância” indica que o método, ao processar um novo registro, leva em consideração as instâncias ou os registros existentes na base de dados. Trata-se de um método de fácil

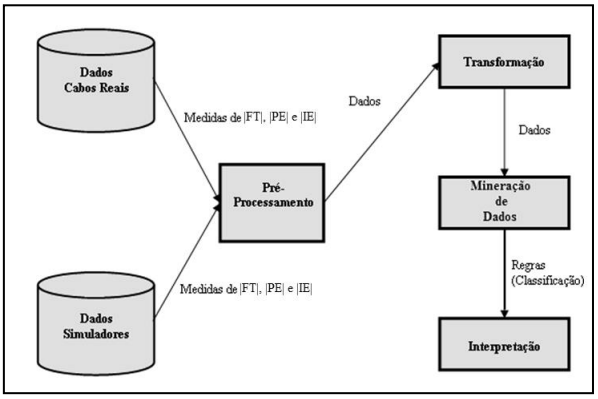
entendimento e implementação e que não requer treinamento prévio para ser aplicado.

3. METODOLOGIA

A metodologia para a proposta dos experimentos neste artigo consiste nas etapas mostradas na Figura 1. A primeira etapa consiste na coleta e seleção de dados obtidos através das medições do módulo da função de transferência |FT|, do módulo do parâmetro de espalhamento S_{11} |PE| e da impedância de entrada |IE|, a partir das bases de dados de medições do enlace local do assinante compostas por cabos reais e simuladores de linha.

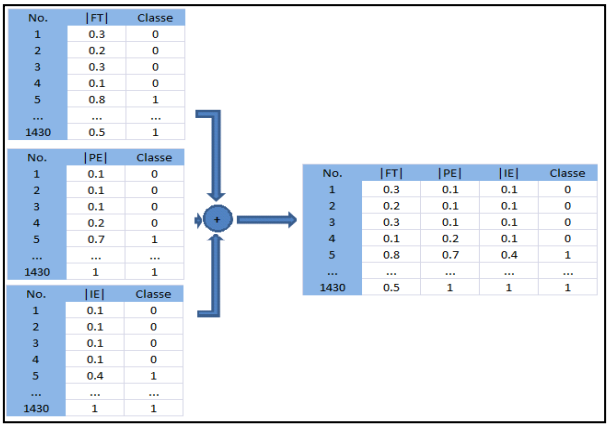
Na segunda etapa, o pré-processamento dos dados, consiste de um tratamento estatístico dos dados selecionados na etapa anterior, com a intenção de se identificar os erros de natureza sistemáticas e/ou aleatórios cometidos nos processos de medição. Para essa etapa utilizou-se o teste de Dixon para detecção de dados (amostras) que não possuem características semelhantes ao restante da população no qual são chamados de *outliers*. A escolha desse teste foi devido à sua facilidade e por se tratar de um método simples e eficaz. Além disso, o fator determinante na sua escolha é: a possibilidade do teste ser aplicado em uma quantidade de amostras considerável. Outra vantagem é que não é necessário o conhecimento a priori da estimativa desses valores (BRITO, 2007).

Figure 1. Diagrama da Implementação da Metodologia



A terceira etapa consiste na transformação dos dados originais obtidos através da primeira etapa, tais dados foram trabalhados de forma, que os mesmos possam facilmente ser utilizados na construção dos classificadores. A Figura 2 mostra a composição da base de dados do enlace local do assinante, composta pelos valores obtidos em medições dos parâmetros citados acima. A base de dados é formada por 1430 instâncias e quatro atributos (incluindo a classe 0 sem extensão ou 1 com extensão) do tipo numérico.

Figura 2. Composição da Base de Dados de Medições.



Na quarta etapa utiliza-se a Mineração de Dados que consiste na construção dos classificadores. Para essa etapa utilizou-se o pacote de mineração de dados Waikato *Environment for Knowledge Analysis* – WEKA (WEKA, 2017). Para o treino, validação e teste dos classificadores binários às seguintes sequências de passos foram feitas:

1-Dividiu-se o banco de dados em três conjuntos: conjunto de treinamento (dados de treinamento com 930 instâncias), conjunto de validação (dados de validação ou amostra de validação 300 instâncias) e o conjunto de teste (dados de teste ou amostra de teste com 200 instâncias).

2-Treinamento (ou aprendizagem): construção de um modelo (classificador) analisando as amostras do conjunto de treinamento.

3-Validação do modelo: aplicação do modelo sobre o conjunto de validação e teste. Calcula-se então, o percentual de acerto das classes preditas pelo modelo em relação às classes esperadas (ou desconhecidas). Esse percentual é chamado de precisão do modelo para o conjunto de validação e teste em questão.

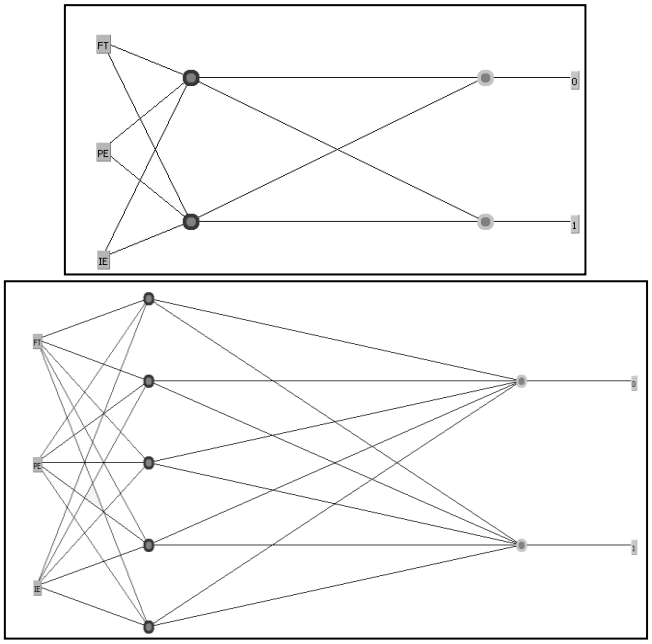
A última etapa consiste na interpretação das regras de classificação obtidas nos experimentos propostos através da obtenção dos valores da acurácia dos modelos e do parâmetro estatístico: erro quadrático médio.

4. OS EXPERIMENTOS

4.1. Classificador Binário com Redes Neurais Artificiais (RNA)

O experimento do classificador binário RNA nos dados do enlace local do assinante, consistiu em avaliar diferentes topologias de RNA com duas ou cinco camadas intermediaria como mostrado na Figura 3(a) e 3(b). Também se avaliou e a influência da taxa de aprendizagem no treinamento da RNA, variando a taxa de aprendizagem de 0.001 a 1.0.

Figura 3. (a) RNA com duas camadas intermediárias. (b) RNA com cinco camadas intermediárias.



A Tabela 1 contém os dados usados para a construção da RNA. Para o treinamento, validação e teste da RNA, admitiram-se um erro mínimo desejado entre a resposta da RNA e a resposta desejada de 0.08. Para cada treinamento, utilizou-se 1000 épocas para treinamento. Os resultados dos erros quadráticos médios (MSE - *Mean Squared Error*) para treino, validação e teste para cada topologia de RNA são mostrados nesta tabela.

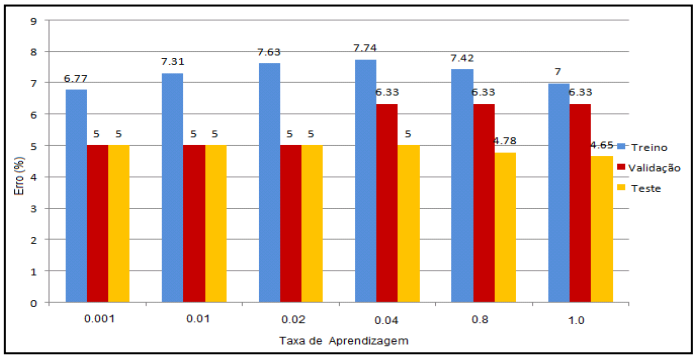
Tabela 1. RNA com duas e cinco camada intermediaria e taxas de aprendizagem.

Topologia da RNA	Nº de Camadas Intermediárias	Taxa de Aprendizagem	MSE de Train.	MSE de Valid.	MSE de Test.
1	2	0.001	0.05322	0.04519	0.03301
2	2	0.01	0.05280	0.04422	0.02924
3	2	0.02	0.05253	0.04410	0.02845
4	2	0.04	0.05189	0.04384	0.02253
5	2	0.8	0.05225	0.04456	0.01290
6	2	1.0	0.05175	0.04460	0.01256
7	5	0.001	0.05225	0.04443	0.03115
8	5	0.04	0.05212	0.03980	0.03062
9	5	0.8	0.05475	0.02647	0.01747
10	5	1.0	0.05345	0.03798	0.02869

Com o acréscimo de mais três camadas totalizando cinco camadas intermediarias como no caso das topologias 7 a 10 da RNA, houve-se uma pequena mudança no poder de computacional da RNA no processo de classificação. Porém, para todas as topologias utilizadas, obteve-se um MSE de treinamento bem abaixo do desejado (0.08), e a RNA conseguiu uma generalização satisfatória para os dados de validação e teste.

O gráfico da Figura 4 mostra a evolução de erro do classificador nas etapas de treino, validação e teste da RNA ao longo do processo de aprendizagem para cada topologia testada. Verifica-se que para todos os casos a RNA conseguiu classificar de maneira adequada os padrões que lhe foram apresentados.

Figura 4. Taxa de Erro para do Classificador RNA x Taxa de Aprendizagem da Rede.



4.2. Classificador Binário com Árvore de Decisão

Segundo (MITCHELL, 2001) a árvore de decisão gerada pelo algoritmo J4.8 é baseada no princípio que o algoritmo tende a gerar árvores menores (teoria mais simples) a partir das instâncias de treinamento, utilizando a heurística da entropia.

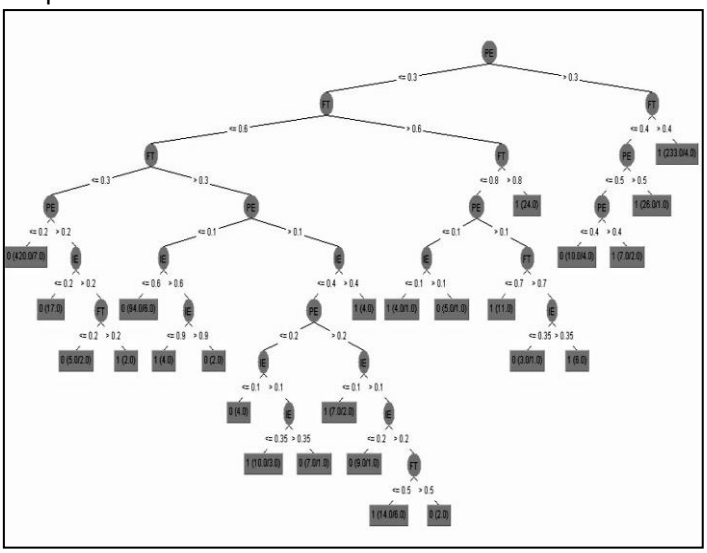
A Tabela 2 contém os dados usados para a construção da Árvore de Decisão utilizando a base de dados de medições do enlace do assinante. Para as etapas de treinamento, validação e teste utilizaram-se alguns valores para os parâmetros: fator de confiança e para o número mínimo de objetos por folha. Os resultados dos (MSE) para treino, validação e teste também são mostrados nesta tabela.

Tabela 2. Configuração da Árvore de Decisão.

Fator de Confiança (C)	Nº de Objetos Minímos (M)	MSE de Train.	MSE de Valid.	MSE de Test.
0.15	2	0.0701	0.0590	0.0468
0.35	2	0.0570	0.0402	0.0059
0.55	2	0.0541	0.0400	0.0005
0.15	3	0.0713	0.0590	0.0468
0.35	3	0.0610	0.0450	0.0059
0.55	3	0.0558	0.0420	0.0055
0.15	5	0.0727	0.0590	0.0468
0.35	5	0.0690	0.0463	0.0081
0.55	5	0.0633	0.0390	0.0076

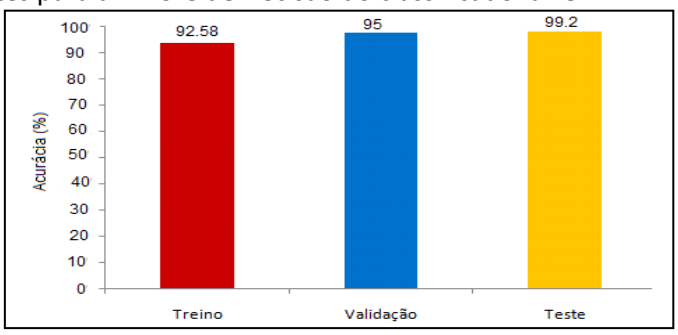
No que diz respeito aos valores obtidos na Tabela 2, percebe-se que o melhor resultado segundo o MSE de treinamento, validação e teste foi alcançado com os valores de 0.55 para o fator de confiança e 2 para o número mínimo de objetos por folha. Neste caso, optou-se por esses valores de parâmetros para a construção do classificador binário J4.8. Para validação e estimação da acurácia de testes, foi utilizado o método 10 *fold Cross-Validation,* sendo gerada a árvore de decisão apresentada na Figura 5.

Figura 5. Árvore de Decisão para o classificador J4.8.



Os valores da acurácia (percentual de classes preditas corretamente) obtidos na utilização da árvore de decisão J4.8 estão apresentados na Figura 6. Tais resultados expressão o desempenho do classificador binário para as etapas: treino, validação e teste. Como pode ser observado nas estatísticas do classificador avaliado, houve uma melhora que diz à acurácia de teste.

Figura 6. Acurácia de Testes para a Árvore de Decisão do classificador J4.8.



4.3. Classificador Binário com *K-NN*

O classificador *K-NN* consiste no aprendizado baseado na analogia dos elementos da base de dados. Para o conjunto de treinamento cada elemento deste conjunto representa um ponto no espaço n-dimensional (BOSCARIOLI; TABUSADANI; BIDARRA, 2008).

A metodologia desse classificador para determinar a classe de um elemento que não pertença ao conjunto de treinamento é a seguinte: o classificador *K-NN* procura *K* elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância. Estes *K* elementos são chamados de *K*-vizinhos mais próximos. Verifica-se quais são as classes desses *K* vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido.

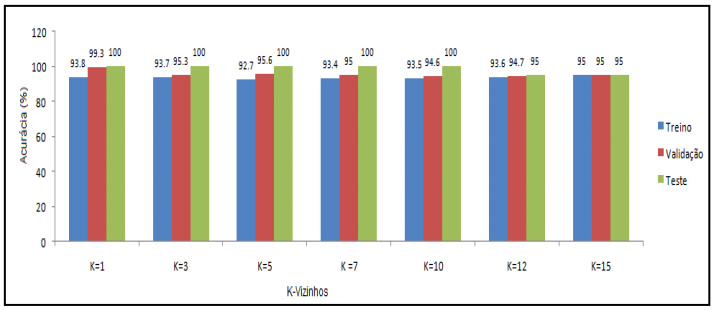
K-NN é um classificador que possui apenas um parâmetro livre (o número de *K*-vizinhos), que é controlado pelo usuário com o objetivo de obter uma melhor classificação. Sendo assim, para os experimentos neste trabalho obteve-se a Tabela 3 com a variação do número de *K*-vizinhos (1 a 15 vizinhos mais próximos) e os valores obtidos para MSE em relação às etapas de treino, validação e teste do classificador binário com *K-NN*.

Tabela 3. Variação do *K-NN* e o MSE do Classificador.

K - Vizinhos	MSE de Train.	MSE de Valid.	MSE de Test.
1	0.0520	0.0113	0.0041
3	0.0471	0.0273	0.0041
5	0.0526	0.0350	0.0056
7	0.0499	0.0372	0.0057
10	0.0513	0.0356	0.0088
12	0.0497	0.0421	0.0221
15	0.0528	0.0423	0.0213

Em relação aos valores obtidos na Tabela 3 percebe-se que os resultados segundo o MSE de treinamento, validação e teste, que as estatísticas dos classificadores avaliados apresentaram poucas diferenças no que diz respeito ao erro médio quadrático, consequentemente os resultados preditos da acurácia devem variar muito pouco. A Figura 7 compara os resultados obtidos variando o valor de K a partir desta métrica. Observa-se que a acurácia em relação às etapas de treino, validação e teste comprova os resultados obtidos na Tabela 3.

Figura 7. Acurácia de Testes para o classificador *K-NN*.



4.4. Classificador Binário com *Naive Bayes*

O classificador binário com *Naive Bayes* é fundamentado no teorema de Bayes da probabilidade condicional. Este classificador calcula a probabilidade de uma instância pertencer a cada uma das classes pré-determinadas, assumindo que não há independência entre os atributos que descrevem a instância (DOMINGOS; PAZZANI, 1997).

A partir desta probabilidade foi possível gerar um modelo de conhecimento com uma regra de decisão que sempre dá como resposta a classe que obteve maior probabilidade após a aplicação do teorema de Bayes (MAP, *Maximum a Posteriori*).

O classificador binário *Naive Bayes* foi induzido a partir do conjunto de treinamento apresentado, utilizando para a validação e estimação da acurácia de teste o método 10 *fold Cross-Validation*. Os valores estatísticos do MSE obtidos da utilização deste classificador estão apresentados na Tabela 4.

Tabela 4. Estatísticas do Classificador Binário Naive Bayes.

	Treino	Validação	Teste
Instâncias Corretamente Classificadas	91.2%	94 %	92 %
Instâncias Erroneamente Classificadas	8.8 %	6 %	8 %
Erro Quadrático Médio (MSE)	0.0691	0.0511	0.0621

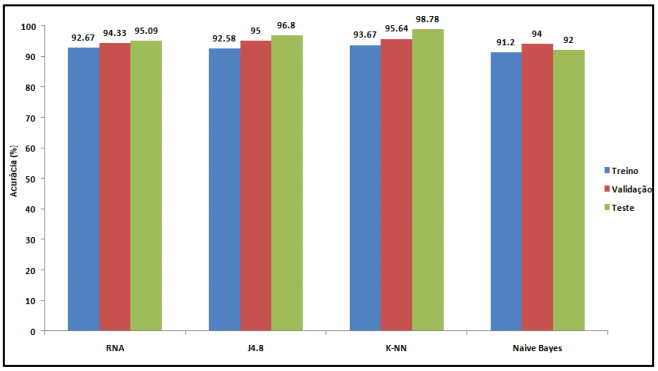
Pelos valores obtidos nos experimentos, segundo a Tabela 4, percebe-se que devido ao fato deste classificador fazer a suposição que os valores das características são independentes dentro da classe, o mesmo apresentou um bom desempenho de predição apesar da sua premissa ingênua e simplista.

5. ANÁLISE DOS RESULTADOS

Os experimentos realizados neste trabalho podem ser analisados utilizando o valor médio da acurácia dos classificadores RNA, J4.8 e *K-NN*, isto se dá devido a utilização de diferentes valores de parâmetros, para melhor obtenção da acurácia dos modelos. Porém, para o classificador binário *Naive Bayes* não foi utilizada essa média pelo fato do mesmo não oferecer possibilidade de escolha de um parâmetro adequado.

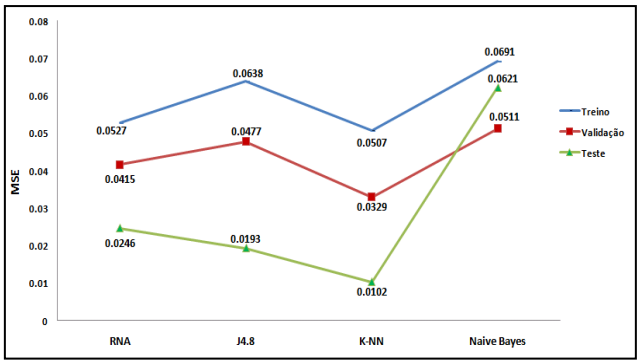
Analisando as estatísticas dos classificadores observa-se que houve pouca diferença no que diz respeito à acurácia do treino, validação e teste. A Figura 8 compara os classificadores nesta métrica.

Figura 8. Acurácia de Testes dos Diferentes Classificadores.



Considerando os valores obtidos do erro quadrático médio MSE para os classificadores utilizados nos experimentos, pode-se avaliar a relação desse parâmetro estatístico com os valores das acurácias dos respectivos classificadores. Através da Figura 9 percebe-se que o classificador *K-NN* possui nas etapas de treino, validação e teste os menores resultados para o erro MSE, comprovando assim os resultados obtidos em relação à acurácia do mesmo, segundo o gráfico apresentado na Figura 8. Porém, como os valores em geral para todos os classificadores utilizados nos experimentos, não apresentam uma discordância tão alta nos seus valores, qualquer um dos classificadores tendem a ter um bom desempenho na classificação de um enlace local do assinante com presença ou não de extensões.

Figura 9. Comparativo das Estatísticas do MSE dos Classificadores.



6. CONSIDERAÇÕES FINAIS

Neste trabalho foram propostos alguns experimentos de extração de conhecimento utilizando métodos baseados em Mineração de Dados. Estes experimentos servem para adicionar conhecimentos relevantes aplicados ao problema de detecção de extensões em determinados enlaces do assinante. Nesse contexto, destaca-se a necessidade de conhecer a infraestrutura do enlace do local para que a mesma possa transportar um serviço final com qualidade ao assinante.

Dentro da proposta deste trabalho ressaltou-se que as informações contidas na base de dados das operadoras, quando obtidas de forma confiáveis, permitem que a tarefa de qualificação do enlace seja feita através das medições dos parâmetros físicos. Diante desse fato, a contribuição desse trabalho se deu na descoberta de conhecimento contido em tais medições. Os resultados obtidos para cada experimento foram comprovados pela obtenção da acurácia, e do parâmetro estatístico MSE de cada modelo obtido. Dentre esses resultados alguns foram obtidos variando os parâmetros importantes dos classificadores escolhidos.

Neste ponto é relevante dizer que todos os classificadores obtidos neste trabalho quando utilizados de forma adequados, fornecem resultados confiáveis para tomada de decisões pelas operadoras de serviços DSL. Porém, do ponto de vista mais eficiente, em relação a sua estrutura e escolha dos seus parâmetros adequados, a Rede Neural é a que permite melhor utilização da descoberta de detecções de extensões no cenário de estudo deste trabalho. Em relação a trabalhos futuros tem-se a possibilidade de combinar mais de uma técnica. Desse modo se fará a utilização de técnicas de inteligência computacional híbrida melhorando ainda mais o desempenho dos experimentos.

REFERÊNCIAS

BOSCARIOLI, C.; TABUSADANI, F. Y.; BIDARRA, J. O uso Integrado de K-NN e Scatter Plots 2D na mineração visual de dados. Universidade Estadual do Oeste do Paraná, 2008.

BRITO, E. Metodologia para a medição de parâmetros relacionados com a qualificação do enlace digital do assinante. Belém: Universidade Federal do Pará, 2007.

DOMINGOS, P; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, v.29, p.103-137, 1997. https://doi.org/10.1023/A:1007413511361

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining, Menlo Park: AAAI Press, 1996.

GOLDSCHMIDT, R.; PASSOS, E. Data Mining um guia prático. Rio de Janeiro: Elsevier, 2017.

MITCHELL, T.M. Machine learning. McGraw-Hill, 2001.

MORIK, K. The Representation Race – preprocessing for handling time phenomena. In: European Conference on Machine Learning 2001, Lecture Notes in Artificial Intelligence 1810. Proceedings… Berlin: Springer Verlag, 2001.

MUNCINELLI, G. Qualificação de linha para serviço ADSL. In: CININTEL 2001, CONGRESSO INTERNACIONAL DE INFRAESTRUTURA PARA TELECOMUNICAÇÕES, 4. Anais... 2001.

WEKA. 2017. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka.> Acesso em: jun. 2017.