PROJETO DE APRENDIZADO DE MÁQUINA - GRUPO 2

Manoela C. B. Silva, Randal Gasparini, Vinicius A. Reis
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar) - Campus Sorocaba
18052-780, Sorocaba, São Paulo, Brasil
{*manoelacamila.silva, gasparini.randal, angiolucci*} *gmail.com*

***Resumo*—Aprendizado de máquina (AM) é um dos ramos com maior destaque na área de estudos sobre inteligência artificial atualmente. É feito um grande esforço nesta área para a criação e melhoria de algoritmos que conferem aos programas a capacidade de aprender. No presente trabalho são aplicados algoritmos clássicos de *classificação* em uma base de dados conhecida, com o intuito de analisar o desempenho destes algoritmos. Juntamente com a comparação de desempenho em função da *acurácia média* obtida em cada método, são propostas melhorias no tratamento de dados e ajustes individuais em cada método que possam melhorar o desempenho de classificação.**

***Palavras-Chave*-aprendizado de máquina; classificação; regressão logística; redes neurais; SVM; k-NN;**

I. INTRODUÇÃO

Aprendizado de máquina é atualmente um dos ramos da área de inteligência artificial com maior velocidade de desenvolvimento [1]. Na literatura não é incomum autores que afirmam que para uma máquina ser considerada inteligente de fato, deve ser capaz de aprender [2].

Sendo assim, os métodos de aprendizado de máquina têm ganhado cada vez mais notoriedade no campo da inteligência artificial pois são capazes, através de diferentes técnicas, de *classificar*, *agrupar* e *estimar* características para os mais variados problemas, situações e conjuntos de objetos do mundo real. A atenção crescente por parte da comunidade científica e o constante incremento de poder computacional têm permitido a criação e o aprimoramento de técnicas de aprendizado de máquina cada vez mais robustas.

O presente trabalho tem como objetivo a implementação e análise de desempenho de quatro classificadores bem conhecidos na literatura, exibidos na Tabela I. Cada classificador é executado sobre uma base de dados pública, previamente compilada e tratada, retirada de um censo indonésio, realizado com mulheres casadas, que relaciona aspectos socioeconômicos com a preferência por uma entre três opções de uso de métodos contraceptivos [3].

Este documento se encontra estruturado como segue: na Seção II são descritas de forma detalhada a estrutura e características da base de dados utilizada, bem como todo o pré-processamento realizado sobre a mesma. Essa etapa visa a minimização da influência dos dados sobre o desempenho do classificador [8].

Tabela I
CLASSIFICADORES IMPLEMENTADOS

Classificadores
k-vizinhos mais próximos (k-NN) [4]
Regressão Logística com Regularização [5]
Redes Neurais com *backpropagation* [6]
Máquinas de Vetores de Suporte (SVM) [7]

A metodologia experimental é descrita com detalhes na Seção III, e descreve os modelos e etapas executados para obtenção de parâmetros específicos para cada algoritmo de classificação, assim como seus processo de validação e teste. Apresentando também a definição das medidas de desempenho utilizadas para análise.

Na Seção IV são apresentados os resultados finais obtidos, e uma análise comparativa dos classificadores. O desempenho de cada técnica é analisado, principalmente, em termos de sua acurácia média. Métodos mais robustos como Regressão Logística e Máquinas de Vetores de Suporte apresentam um desempenho nitidamente superior, o qual é discutido levando-se em consideração a quantidade dos atributos de cada amostra na base de dados.

Por fim, na Seção V são apresentadas as principais conclusões obtidas da execução desse trabalho.

II. BASE DE DADOS

A base de dados utilizada foi criada por Tjen-Sien Lim [3] e possui acesso público[1], consistindo de um subconjunto dos dados obtidos através de um censo indonésio, realizado em 1987 com mulheres casadas, com o intuito de verificar a escolha de método contraceptivo utilizado pelas mesmas.

A base em questão é composta de 1473 amostras contendo 9 atributos cada e não possui nenhum valor ausente. Cada um dos atributos corresponde à resposta de uma pergunta, realizada a uma das mulheres entrevistadas. A questão a qual cada atributo se refere, assim como o tipo de dado e os valores permitidos ao mesmo, podem ser visualizados na Tabela II. Também é possível notar que a maioria dos atributos da base são do tipo categórico.

[1] Contraceptive Method Choice Data Set. Disponível em: https://goo.gl/fPIfiJ. Acessado em 19/05/2015.

Tabela II
CARACTERÍSTICA DOS ATRIBUTOS DA BASE DE DADOS

Pergunta	Tipo	Valores
Idade da esposa	Númerico	> 0
Grau de instrução da esposa	Categórico Ordinal	1-baixo, 2, 3, 4-alto
Grau de instrução do esposo	Categórico Ordinal	1-baixo, 2, 3, 4-alto
Número de filhos já nascidos	Númerico	$>= 0$
Esposa segue ao Islam?	Binário	1-Sim, 0-Nao
Esposa atualmente trabalha?	Binário	0-Sim, 1-Nao
Ocupação do esposo	Categórico Nominal	1, 2, 3, 4
Padrão de Vida	Categórico Ordinal	1=baixo, 2, 3, 4=alto
Exposicao publica	Binário	0-Boa, 1-Ruim
Método contraceptivo usado	Classe de Saída	1-Nenhum, 2-Longo prazo, 3-Curto prazo

A. *Pré-Processamento*

Na etapa de pré-processamento, os dados foram inicialmente analisados a fim de encontrar e eliminar inconsitências e redundâncias.

Após a análise, foram eliminadas da base 48 amostras redundates, isto é, tuplas que apareciam mais de uma vez, mantendo apenas uma única instância de cada objeto, a fim de não influenciar o classificador em relação àquela amostra. Também foram eliminadas da base 129 amostras inconsistentes, que representavam conjuntos de tuplas com valores de atributos idênticos, entretanto com rótulos divergentes. A decisão de eliminá-las deve-se ao fato de que, como não era possível determinar qual dos rótulos era o correto, as mesmas foram consideradas não confiáveis para o aprendizado dos classificadores, podendo influenciar o processo de forma negativa.

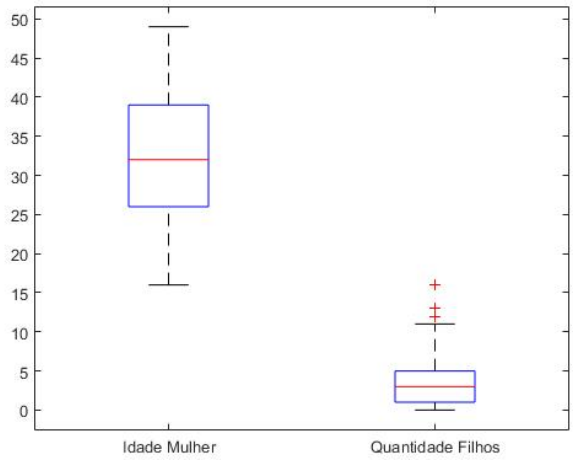

Figura 1. Boxplot - Análise de *outliers*

Por fim, foi realizado o *boxplot* dos dados numéricos para análise da existência de possíveis *outliers*. Foram encontrados alguns *outliers* no atributo referente ao número de filhos já nascidos, como pode ser visto na Figura 1. Como o número de *outliers* encontrados foi relativamente baixo (7), e os mesmos são valores possíveis considerando o cenário em questão, optou-se por mantê-los na base, uma vez que podem ajudar na precisão de ajuste da hipótese dos classificadores.

A base final utilizada nos experimentos não contém redundâncias, nem inconsistências e possui 1296 amostras embaralhadas. A fim de realizar uma análise visual dos dados restantes em tal base, foi utilizado o método do PCA [9], através do qual foram obtidos os três principais atributos dentre os dados e feito um plot 3D da base (Figura 2).

Outra situação abordada foi o embaralhamento da base, haja visto que os dados estavam concentrados em grupos de rótulos iguais. Tal ação se fez necessária para garantir que, todas as partições em que as amostras foram divididas, obtivessem uma heterogeneidade em relação aos rotulos.

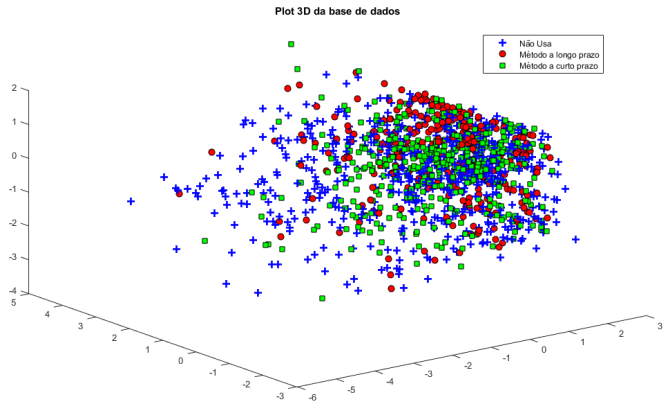

Figura 2. Plot em 3D da base final

III. METODOLOGIA EXPERIMENTAL

A metodologia utilizada para análise e preparo dos metódos classificatórios implementados no presente trabalho é composta de três fases, sendo elas: escolha de parâmetros, teste do modelo escolhido e treinamento do classificador.

A fase de escolha de parâmetros consiste em ajustar os parâmetros utilizados pelos classificadores, de forma a maximizar o seu desempenho. Para tanto, foi adotada a busca em *grid* (*grid search*), a fim de verificar o desempenho obtido por cada classificador, considerando diversas combinações de valores para seus parâmetros.

Para as técnicas de regressão logística e SVM, tal busca foi realizada utilizando a metodologia de *cross-validation* com *10-fold*. A escolha do *cross-validation*, nesse caso, deu-se em relação ao fato de ser um modelo exaustivo e que possui uma melhor precisão, quando comparado com o *holdout*. Mas, principalmente, por se tratar de uma base

relativamente pequena, onde tal procedimento se tornou viável no tempo disponível. Além disso, dessa forma, não foi necessário fazer uma validação posterior do conjunto de valores selecionados, uma vez que a mesma foi realizada ao mesmo tempo em que os valores foram testados.

Para a técnica de redes neurais, uma vez que a mesma exige um maior poder e tempo de processamento para ser executada, a busca em *grid* foi realizada com *holdout* 80/20 e, posteriormente, na etapa de teste do modelo escolhido, foi então executado o *10-fold* para os parâmetros selecionados.

Os intervalos adotados para cada parâmetro, assim como os seus passos, foram setados como segue:

- O valor de λ para os métodos de regressão logística e rede neural, e os valores de C e γ para o SVM, foram testados no intervalo $[10^{-3}, 10^{3}]$, utilizando como passo o incremento de 1 na potência;
- O grau do polinômio da regressão logística foi testado no intervalo $[0, 3]$, com passo 1;
- O número de camadas intermediárias foi testado com 1 e 2.
- O número de neurônios em cada camada intermediária da rede neural foi testado no intervalo $[$*input*$+1,$ *input*$+4$, com passo 1, onde "*input*"é a dimensão da camada de entrada;
- O tipo do kernel do SVM: Linear, Radial, Sigmoidal ou Polinomial.

O método classificatório do k-NN utilizou-se de uma busca em *grid* mais simples, uma vez que seu único parâmetro é o k (número de vizinhos). Para esse último método, então, foram testados valores de ks variando de 1 até 9, com passo igual a 1, através da aplicação das metodologias *holdout* 80/20 e *10-fold*. Ao fim, o k com melhor desempennho obtido foi o escolhido para execução do classificador.

A Tabela III apresenta os valores escolhidos para cada parâmetro.

Tabela III
VALORES DOS PARÂMETROS

Classificadores	Parâmetros	Valores Selecionados
Regressão Logística	λ	1
	grau polinômio	0
Redes Neurais	λ	0.001
	qntd camadas int.	1
	neurônios camada int.	11
SVM	C	1000
	γ	0.001
	tipo kernel	Radial
k-NN	k	4

A etapa de teste consistiu na execução das tarefas de treino e teste do classificador, com os parâmetros selecionados previamente, a fim de avaliar o desempenho do mesmo, assim como possíveis *underfitting* e *overfitting*. Nessa etapa, principalmente devido ao tempo disponível para execução dos experimentos, optou-se por uma única execução da validação cruzada com 10 partições, que, com exceção do método de rede neural, que utilizou *holdout* 80/20 na seleção de parâmetros, foi feita durante a execução da busca em *grid*.

Para medir o desempenho obtido por cada um dos classificadores foram utilizadas medidas de desempenho largamente conhecidas na literatura, sendo elas:

- Acurácia(%): amostras devidamente classificadas;
- Precisão(%): capacidade do classificador em evitar predições equivocadas em relação a cada classe;
- F-medida: média harmônica entre precisão e sensitividade.

Além disso, foram construídas as curvas de aprendizado de cada método, com exeção do k-NN, utilizando 90% da base como conjunto de treino e os 10% restantes como teste, para análise de *underfitting* e *overfitting*. Os pontos das curvas foram dados iniciando com 10% do conjunto de treinamento e, a cada passo, incrementando-o em 10%, até que 100% da parcela da base reservada para treinamento fosse utilizada. Para o k-NN, entretanto, foi construída a curva de acurácia em relação a variação de k, para análise visual dos desempenhos obtidos.

Por último, foi realizada a etapa de treinamento final dos classificadores, tornando-os aptos para a classificação de amostras desconhecidas. Nessa fase, foi realizado um último treinamento, utilizando os parâmetros previamente selecionados e validados, e a base completa de amostras conhecidas.

Vale ressaltar que, tratando-se de um problema multiclasses, para o método classificatório de regressão logística, foi utilizada a abordagem *One vs. All*[2] para treinamento do classificador. A abordagem também foi utilizada para obtenção das medidas de desempenho em todos os classificadores analisados. Também é importante dizer que, a técnica de SVM foi implementada utilizando a biblioteca externa LIBSVM [10].

IV. RESULTADOS

Nessa seção são exibidos os resultados obtidos da execução de cada um dos métodos classificatórios analisados no presente trabalho. Além disso, também é feita uma análise comparativa do desempenho de cada um dos métodos em relação a base utilizada.

Numa primeira etapa da análise, foram abordados os resultados obtidos através da execução do *10-fold*, com os melhores parâmetros obtidos na realização da busca em *grid*. A Tabela IV contém os valores médios de acurácia, precisão e f-medida de cada um dos classificadores, juntamente com seu desvio padrão; tais valores encontram-se ordenados pela acurácia.

[2]Vídeo-aula disponível em: https://class.coursera.org/ml-005/lecture/38. Acessado em: 05/06/2015.

Tabela IV
RESULTADOS OBTIDOS PELOS MÉTODOS CLASSIFICATÓRIOS DURANTE A EXECUÇÃO DO CROSS-VALIDATION COM 10-FOLD

	Acurácia (%)	**Precisão (%)**	**F-medida**
Regressão Logística	71.45 ± 3.38	55.21 ± 6.20	0.53 ± 0.05
SVM - Radial	71.04 ± 2.44	55.19 ± 5.45	0.52 ± 0.04
SVM - Linear	68.31 ± 3.21	51.70 ± 7.31	0.48 ± 0.05
SVM - Sigmoidal	68.31 ± 3.21	51.67 ± 7.31	0.48 ± 0.05
SVM - Polinomial	68.16 ± 3.17	50.70 ± 6.04	0.49 ± 0.05
Redes Neurais	64.30 ± 2.33	64.15 ± 2.53	0.34 ± 0.03
k-NN	60.13 ± 2.88	35.56 ± 4.91	0.37 ± 0.04

Como pode-se notar, os classificadores que utilizam as técnicas de SVM e Regressão logística foram os que obtiveram um melhor desempenho, em relação a acurácia na classificação correta das amostras, durante a etapa de validação e testes do mesmo, sendo a diferença de desempenho entre eles relativamente baixa.

Analisando os resultados obtidos pela técnica de SVM, com os diversos tipos de *kernels*, constatatou-se aquilo que era já esperado, segundo dados da literatura, que, por se tratar de uma base onde existem muito mais amostras do que atributos, e o número de amostras não chega a atingir uma magnitude exorbitante, o SVM com *kernel* radial obteria um melhor resultado, quando comparado com os demais *kernels*.

Também pode-se notar que o método do k-NN foi o que obteve o pior desempenho, confirmando o que foi exposto em [3]. Tendo em vista que tal método é extremamente sensível a distâncias, pode-se atribuir tal resultado ao pouco espalhamento das amostras da base utilizada. Como pode ser visto na Figura 2, a distância e heterogeneidade entre as diversas amostras da base são consideravelmente baixas.

Quando analisados os resultados obtidos pelos diversos classificadores de forma comparativa, notou-se que, o método de redes neurais, apesar de ter obtido um desempenho consideravelmente inferior em relação aos métodos de regressão logística e SVM, foi o qual obteve uma melhor taxa de precisão. Contudo, tratando-se de um cenário com pouca rigorosidade em relação à proximidade dos resultados, e onde a acurácia é a medida mais importante, o desempenho de tal método é considerado insatisfatório para o problema abordado.

Também verificou-se que, apesar das técnicas de regressão logística e SVM terem obtidos desempenhos muito próximos, analisando o conjunto de medidas como um todo, o metódo de regressão logística foi o que obteve o melhor desempenho geral.

Após a realização dessa análise preliminar dos resultados, foram então construídas as curvas de aprendizado para os métodos que possuem um processo de treinamento, a fim de analisar a evolução de tal processo e identificar possíveis *underfitting* e *overfitting*, ajustando os parâmetros dos classificadores sempre que notou-se necessário, a fim de eliminar problemas de super ou sub-ajustamento.

Para o método do k-NN, não foram construídas curvas de aprendizado, uma vez que o mesmo não possui um processo de treinamento com obtenção de hipóteses diferenciadas em relação ao tamanho da base. Sendo assim, para esse método, foi construída a curva de acurácia em relação a variação do valor de k, a fim de apresentar uma justificativa visual da escolha de k apresentada na Tabela IV. Tal curva é apresentada na Figura 3.

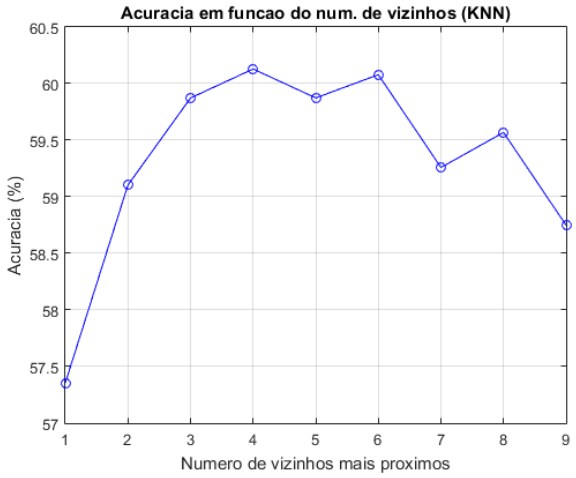

Figura 3. Curvas de Acurácia em relação a K - k-NN

Na Figura 4 são apresentadas as curvas de aprendizado obtidas da execução do método de regressão logística. Aquelas representadas por linhas contínuas representam as curvas de aprendizado obtidas da execução realizada com os parâmetros encontrados pelo *grid*. Como pode-se observar, um alto viés foi verificado, e as taxas de erro tiveram uma variação muito baixa entre o ínicio das curvas e o fim das mesmas, apontando um aprendizado quase que completamente realizado com um pequeno número de amostras e um comportamento tendencioso por parte do classificador.

Com o intuito de melhorar o processo de aprendizado de tal classificador, utilizou-se a técnica de diminuição do valor de λ e aumento da quantidade de atributos (a partir do aumento no grau dos polinômios gerados), para que se fosse obtido uma diminuição do viés. Para tanto, alguns valores foram testados de modo empírico e aqueles que obtiveram melhores resultados foram $\lambda = 0.05$ e grau do polinômio = 2. As curvas obtidas do ajustamento desses parâmetros são aquelas representadas pelas linhas tracejadas.

Na construção das curvas do método SVM, também foi possível notar o mesmo comportamento obtido inicialmente pela regressão logística, quando executado com os parâmetros obtidos pelo *grid*, como pode ser observado na Figura 5. Nesse caso, também com o intuito de diminuir o alto viés apresentado, foi aumentado o valor de C para 100.000, o que ocasionou nas curvas representadas pelas

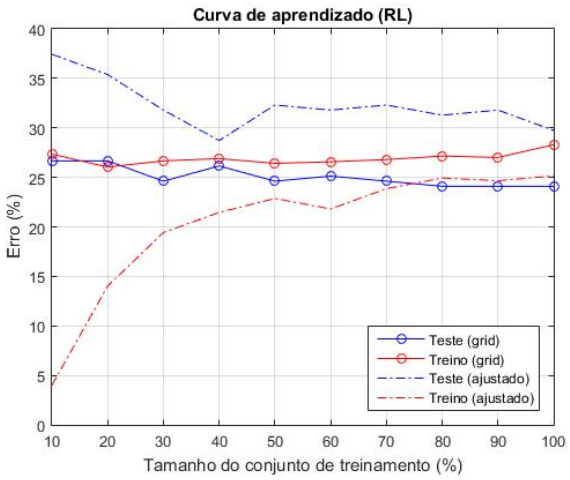

Figura 4. Curvas de aprendizado - Regressão Logística

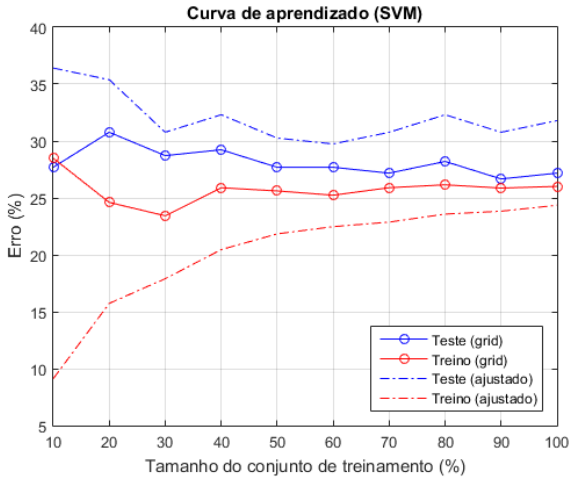

Figura 5. Curvas de aprendizado - SVM com kernel radial

linhas tracejadas, que possuem um comportamento mais similar ao esperado.

As curvas de aprendizado do método de redes neurais, por outro lado, quando construídas a partir da execução com os parâmetros obtidos pelo *grid*, apresentaram um comportamento muito próximo do esperado, como pode ser observado na Figura 6. Logo, para essa técnica, nenhum ajuste adicional de parâmetros foi realizado.

A Tabela V apresenta os valores das medidas de desempenho obtidos após os ajustes de parâmetros das técnicas de regressão logística e SVM.

Observa-se uma pequena queda na taxa de acurácia dos métodos após o ajuste de parâmetros. Entretanto, tal queda não influencia significativamente o desemepenho do algoritmo, e possui o benefício de reduzir significativamente os problemas de ajustamento inicialmente apresentados pelos classificadores, expostos nas Figuras 4 e 5. Este resultado justifica a utilização destes parâmetros empíricos no treinamento final do classificador, ainda que com uma pequena perda de desempenho em relação a acurácia.

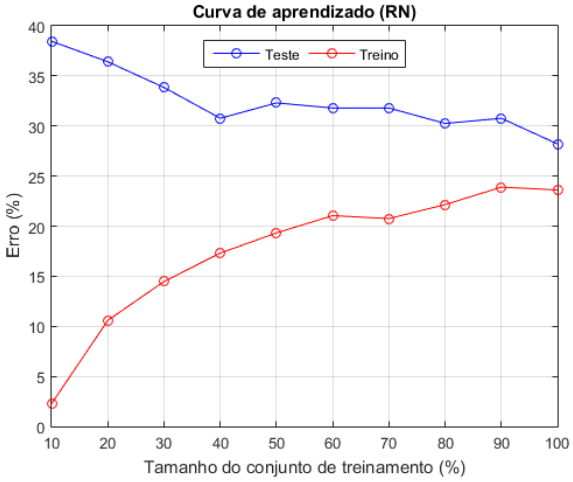

Figura 6. Curvas de aprendizado - Rede Neural

Tabela V
RESULTADOS OBTIDOS PELOS MÉTODOS CLASSIFICATÓRIOS DE REGRESSÃO LOGÍSTICA E SVM - RADIAL APÓS AJUSTE DE PARÂMETROS

	Acurácia (%)	**Precisão (%)**	**F-medida**
Regressão Logística	70.73 ± 3.01	53.84 ± 5.17	0.53 ± 0.05
SVM - Radial	70.68 ± 2.33	53.80 ± 3.32	0.53 ± 0.04

Esta análise sugere que devido ao método de execução exaustivo do *grid* para regressão logística e SVM, foram obtidos parâmetros super-ajustados em relação para estes métodos. Um super-ajustamento causa um aprendizado tendencioso por parte destes classificadores, ocasionando o comportamento observado nas suas curvas de aprendizado iniciais. Por outro lado, a rede neural utiliza uma metodologia menos precisa na sua busca por parâmetros ótimos, o que resultou em valores para estes parâmetros em que o aprendizado ocorreu de modo mais incremental, e melhor distribuído conforme o aumento da quantidade de amostras de treinamento.

Por fim, pode-se constatar que nenhum dos métodos foi capaz de alcançar uma acurácia superior a 70%, o que pode ser justificado pelo fato da base utilizada possuir características e estrutura que a tornam de difícil predição, como exposto em [3]. Sendo assim, os desempenhos finais obtidos pelos classificadores mais acurados podem ser considerados satisfatórios para o cenário em questão.

V. CONCLUSÕES

A execução dos experimentos realizados para a elaboração deste trabalho forneceram um conjunto sólido de conhecimentos sobre o tratamento de dados e o funcionamento de algoritmos de classificação. A maioria das dificuldades

encontradas, como a ocorrência de super-ajustamento, puderam ser solucionadas com técnicas encontradas na própria literatura de aprendizado de máquina.

Em relação aos diferentes algoritmos de classificação analisados, comprovou-se a eficiência de métodos reconhecidamente mais robustos como SVM e Regressão Logística. A natureza dos dados utilizados neste trabalho, como já observada em trabalhos relacionados [3], apresenta um cenário de difícil predição por parte dos algoritmos e impacta com maior intensidade em técnicas baseadas em distâncias, como o k-NN.

De modo geral, este trabalho foi fundamental para a fixação de conhecimento tanto prático como teórico sobre aprendizado de máquina, e seus resultados experimentais vão ao encontro do que está disponível na literatura científica sobre o tema.

REFERÊNCIAS

[1] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1993.

[2] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, pp. 89–109, 2001.

[3] Y.-S. S. Tjen-Sien Lim, Wei-Yin Loh, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classication algorithms," *Machine Learning*, vol. 40, pp. 203–228, 1998.

[4] D. C. S. Robert Meersman, Zahir Tari, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2888.

[5] S. le Cessie and J. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, pp. 191–201, 1992.

[6] S.-C. Wang, *Interdisciplinary Computing in Java Programming - Part II*, ser. The Springer International Series in Engineering and Computer Science. Springer US, 2003, vol. 743.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[8] G. H. de Almeida Prado Alves Batista, "Pré-processamento de dados em aprendizado de máquina supervisionado," Ph.D. dissertation, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2003.

[9] L. I. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, p. 52, 2002.

[10] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

APÊNDICE

Com exceção do método de classificação SVM, todos os métodos utilizam bibliotecas de *toolboxes* padrões do *Matlab/Octave*. Para a correta execução das rotinas que utilizam de alguma forma o SVM, é necessário a instalação prévia da biblioteca *LIBSVM*. A documentação para instalação pode ser encontrada em http://www.csie.ntu.edu.tw/~cjlin/libsvm/

A fim de facilitar a organização e execução dos *scripts*, a implementação do projeto se encontra dividida em subpastas, cada qual correspondente a um método classificador ou funcionalidades utilizadas por todos eles. Desse modo, objetiva-se facilitar a compreensão das rotinas utilizadas.

O *script MenuPrincipal.m*, contido na pasta raiz dos códigos fontes deste projeto, conduz todo o experimento, apresentando opções em forma de menu, por onde é possível visualizar detalhes sobre a base de dados utilizada e acessar opções específicas de cada método de classificação. Este mesmo menu permite a reprodutibilidade dos experimentos reportados neste relatório e apresenta em seu primeiro nível as seguintes opções:

k-NN
: Exibe as opções para execução e análise relacionadas ao método dos k-vizinhos.

Regressão Logística
: Exibe as opções para execução e análise relacionadas ao método de Regressão Logística

Redes Neurais
: Exibe as opções para execução e análise relacionadas ao método de classificação por Redes Neurais com *backpropagation*

SVM
: Exibe as opções para execução e análise relacionadas ao método de classificação Máquinas de Vetores de suporte.

Plotar a base em 2D
: Aplica a técnica de redução de dimensionalidade na base de dados e exibe os dados em um plano cartesiano ortogonal de duas dimensões.

Plotar a base em 3D
: Aplica a técnica de redução de dimensionalidade na base de dados e exibe os dados em um espaço cartesiano ortogonal de três dimensões.

Sair
: Fecha o menu principal e encerra o programa

Para as opções que dão acesso aos métodos de classificação, existe ainda uma segunda hierarquia de menus, por onde é possível executar testes de validação cruzada e exibir gráficos com o comportamento do método selecionado, entre outras opções. É importante notar que cada método pode apresentar opções diferentes entre si. A segunda hierarquia de menus é composta por:

Executar *Holdout* 80/20
: Permite a execução de uma validação do tipo *holdout* sobre a base de dados, exibindo no fim da execução as estatísticas de acurácia, f-medida e precisão obtidas.

Executar *10-fold* sobre a base
: Permite a execução de uma validação do tipo *10-fold cross validation* sobre a base de dados para um determinado método, exibindo no fim da execução as estatísticas de acurácia, f-medida e precisão obtidas.

Executar busca em *grid*
: É executada a busca em *grid* para um determinado método a fim de encontrar os melhores valores de parâmetros para o mesmo, dentro de um intervalo pré-determinado.

Treinar classificador
: Para os métodos que necessitam de treinamento prévio executa a rotina de treinamento do classificador, utilizando toda a base conhecida e os melhores parâmetros encontrados.

Classificar uma nova amostra
: Com base no treinamento realizado faz a classificação de novas amostras fornecidas pelo usuário.

Exibir curva de acuracia em funcao de *k*
: Para o método de classificação k-NN, exibe uma curva com a acurácia obtida em função do valor da quantidade dos *k* vizinhos mais próximos utilizados para a predição.

Exibir curva de aprendizado
: Para os demais métodos de classificação exibe a curva de aprendizado baseada nos parâmetros obtidos na fase de validação cruzada. Para métodos que tiveram parâmetros ajustados manualmente para melhorar seu desempenho, exibe uma curva adicional com a evolução do aprendizado para estes parâmetros.

Retornar ao menu principal
: Volta ao nível superior de hierarquia de menus.

Encerrar programa
: Fecha o menu e encerra o programa

O *script* e a base contidos na pasta 'pre-processamento' não são utilizados durante a execução do programa. Trata-se da base original crua, juntamente com o script que realiza o pré-processamento necessário sobre a mesma. Como a base utilizada já se encontra pré-processada, tal *script* não vem a ser utilizado, porém, é possível ver o seu funcionamento acessando essa pasta e digitando "preProcessamento"no terminal.