



Universidade Federal do Ceará – UFC
Departamento de Computação
Mestrado e Doutorado em Computação

Adonias Caetano de Oliveira

Máquina de Aprendizagem Mínima com Opção de Rejeição

Fortaleza/CE

2016

Adonias Caetano de Oliveira

Máquina de Aprendizagem Mínima com Opção de Rejeição

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação, da Universidade Federal do Ceará, como parte dos requisitos exigidos para obtenção do grau de Mestre em Ciência da Computação.

Prof. Dr. João Fernando Lima Alcântara
Prof. Dr. João Paulo Pordeus Gomes

Fortaleza/CE

2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O45m Oliveira, Adonias Caetano de.

Máquina de Aprendizagem Mínima com Opção de Rejeição / Adonias Caetano de Oliveira. – 2016.
79 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2016.

Orientação: Prof. Dr. João Fernando Lima Alcântara.

Coorientação: Prof. Dr. João Paulo Pordeus Gomes.

1. Máquina de Aprendizagem Mínima. 2. Classificação com Opção de Rejeição. 3. Classificação Binária. I. Título.

CDD 005

Adonias Caetano de Oliveira

Máquina de Aprendizagem Mínima com Opção de Rejeição

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação, da Universidade Federal do Ceará, como parte dos requisitos exigidos para obtenção do grau de Mestre em Ciência da Computação.

Aprovada em

BANCA EXAMINADORA

**Prof. Dr. João Fernando Lima
Alcântara(Orientador)**
Universidade Federal do Ceará

**Prof. Dr. João Paulo Pordeus
Gomes(Coorientador)**
Universidade Federal do Ceará

**Prof. Dr. Amauri Holanda de Souza
Junior**
Instituto Federal de Educação, Ciência e
Tecnologia do Ceará

**Prof. Dr. Ajalmar Rêgo da Rocha
Neto**
Instituto Federal de Educação, Ciência e
Tecnologia do Ceará

Fortaleza/CE

2016

*Este trabalho é dedicado à minha família,
em especial meus pais Valdemir Caetano e Conceição Caetano, irmãos Ananias Caetano e
Mônica de Oliveira e minha namorada Mayara Fernandes*

Agradecimentos

Quero deixar aqui registrado a minha mais sincera gratidão para com todos os que colaboraram, diretamente ou indiretamente, neste trabalho:

A Deus, que me orienta em todos os momentos.

Ao Prof. Dr. João Paulo, por ter me orientado durante este projeto, e mais que meu orientador, foi meu amigo. Obrigado pela confiança em meu trabalho, além da dedicação, paciência e acima de tudo pela orientação de como conduzir este trabalho.

Aos professores participantes da banca examinadora: Prof. Dr. João Paulo Pordeus Gomes, Prof. Dr. João Fernando Lima Alcântara, Prof. Dr. Amauri Holanda de Souza Junior e Prof. Dr. Ajalmar Rêgo da Rocha Neto, pelo tempo dedicado a analisar este trabalho e pelas valiosas colaborações e sugestões.

Aos meus pais Valdemir Caetano e Maria Conceição Caetano, irmã Mônica de Oliveira, irmão Ananias Caetano e namorada Mayara Fernandes pelo apoio incondicional, conselhos, força, amor e amizade sem igual durante todos esses anos. Sem eles nada disto seria possível.

Aos meus amigos Rômulo Lopes, Jonas Rodrigues, Joel Cruz e Mariana Carneiro pelo apoio e amizade durante os anos de pós-graduação.

A todos docentes do Mestrado em Computação (MDCC / UFC) e da Graduação em Ciência da Computação (IFCE Campus Maracanaú), pelo ensino de boa qualidade, atenção e apoio.

À Fundação Cearense de Amparo à pesquisa (FUNCAP) pela concessão de bolsa de estudo como apoio financeiro ao desenvolvimento científico

A todos os colegas da EEEP Gov. Luiz de Gonzaga Fonseca Mota em 2013, de modo especial à diretora Caroline Nepomuceno e coordenadora Helen Raquel ao me liberarem para participar no Processo de seleção e depois cursar o Mestrado em tempo integral.

A todos os colegas da EEEP José de Barcelos, de modo especial à diretora Maria da Glória e coordenadores Paula Parente e Murilo Ribeiro por entenderem a importância e apoiarem a Pós-Graduação. Neste caso, facilitando/condensando os horários de aula do pesquisador para que o desenvolvimento deste trabalho fosse possível.

Ao Núcleo Fortaleza do Projeto Acessibilidade Virtual, através dos amigos Agebson Rocha, Marcos Vinícius, Phyllipe do Carmo, Lívio Siqueira e Francisco Alves.

E a você, leitor, pelo interesse por Aprendizagem de Máquina!

"Assim diz o Senhor: Não se glorie o sábio na sua sabedoria, nem se glorie o forte na sua força; não se glorie o rico nas suas riquezas, Mas o que se gloriar, glorie-se nisto: em me entender e me conhecer, que eu sou o Senhor, que faço beneficência, juízo e justiça na terra; porque destas coisas me agrado, diz o Senhor." Jeremias 9:23,24

Resumo

A Máquina de Aprendizagem Mínima (MLM) é um método de aprendizagem indutivo supervisionado aplicado em problemas de classificação e regressão. Basicamente é um mapeamento entre configurações geométricas dos pontos no espaço de entrada e saída. Com a configuração conhecida para um ponto de entrada no espaço de entrada poderá ser estimada a configuração correspondente no espaço de saída após a obtenção de um modelo de aprendizagem linear simples entre matrizes de distância da entrada e saída. O resultado estimado é então aproveitado para localizar o ponto de saída e, assim, prover uma estimativa para resposta ou indicação da classe.

A MLM tem alcançado um desempenho promissor em vários problemas de classificação e regressão em comparação com outros métodos clássicos de aprendizagem. Entretanto, ainda não foi analisado seu desempenho utilizando estratégia de classificação com opção de rejeição. Essa técnica protege o sistema de apoio à decisão em muitas atividades humanas, sobretudo no domínio da medicina, contra erros excessivos como consequências de tomadas de decisão difíceis. Dessa maneira, potenciais erros são convertidos em rejeição, evitando maior confusão e delegando-os assim para a avaliação de um especialista, ou mesmo, por classificadores mais especializados.

Portanto, a proposta desta dissertação é o desenvolvimento da Máquina de Aprendizagem Mínima (MLM) e suas variantes com opção de Rejeição em problemas de classificação binária, mais especificamente, na classificação de patologias da Coluna Vertebral (PVC-2C), Diabetes (Pima Indians diabetes), sobrevivência do Câncer de Mama (Haberman) e predição de defeitos de software (KC2). A avaliação do desempenho dessas técnicas consiste, em geral, na análise da curva de acurácia-rejeição comparativamente com métodos mais tradicionais de classificação com opção de rejeição que se baseiam os métodos Perceptron Multicamadas (MLP), K-Vizinhos mais Próximos (K-NN) e K-Médias.

Palavras-chaves: Máquina de Aprendizagem Mínima, Classificação com Opção de Rejeição, Classificação Binária.

Abstract

The Minimal Learning Machine (MLM) is an inductive learning method applied to supervised classification and regression problems. It is basically a mapping between points in the geometric configurations of the input and output space. With the known configuration to an entry point in the input space corresponding to the output configuration space after obtaining a simple linear model learning distance between arrays of input and output can be estimated. The estimated result is then passed to locate the exit point and thus provide an estimate for response or indication of the class.

The MLM has reached a promising performance in various classification and regression problems compared with other classical methods of learning. However, it has not yet been analyzed performance using classification strategy of rejection option. This technique protects the system for decision support in many human activities, especially in the field of medicine, against excessive errors like making difficult decision consequences. In this way, potential errors are converted into rejection, avoiding further confusion and delegating them well for the evaluation of an expert, or even, for more specialized classifiers.

Therefore, the purpose of this dissertation is the development of the Minimum Learning Machine (MLM) and its variants with Rejection option in binary classification problems, more specifically, in the classification of diseases Spinal (PVC-2C), Diabetes (Pima Indians diabetes), survival of breast cancer (Haberman) and prediction software defects (KC2). The evaluation of the performance of these techniques is, in general, the analysis of the accuracy-rejection curve compared to more traditional methods of classification with rejection option that are based in methods Perceptron Multilayer (MLP), K-Neighbors More Next (K -NN) and K-Means.

Key-words: Minimal Learning Machine, Classification with Reject Option, Binary Classification.

Lista de ilustrações

Figura 1 – Hierarquia de aprendizado (GAMA; CARVALHO, 2011)	1
Figura 2 – Modelo de Neurônio Artificial de McCulloch e Pitts.	7
Figura 3 – Arquitetura geral de uma rede neural supervisionada (OLIVEIRA, 2001)	8
Figura 4 – Procedimento para obter estimativa.	11
Figura 5 – Classificador Binário com opção de Rejeição. Fonte: (SOUSA; MORA; CARDOSO, 2009a)	18
Figura 6 – Regiões de rejeição e aceitação da regra ótima de reconhecimento. Fonte: (ROCHA NETO, 2011)	20
Figura 7 – Curva de compromisso entre o erro e a rejeição. Fonte: (ROCHA NETO, 2011)	21
Figura 8 – Curva hipotética de compromisso entre o erro e a rejeição.	21
Figura 9 – Variação do tamanho da região de rejeição de acordo com o valor de <i>limiar</i>	24
Figura 10 – Classificação convencional quando <i>limiar</i> = 0.	24
Figura 11 – Taxa de acerto para um certo <i>limiar</i> e <i>n</i> realizações	25
Figura 12 – Taxa de rejeição para um certo <i>limiar</i> e <i>n</i> realizações	25
Figura 13 – Exemplo de escolha de melhor desempenho quando $Wr = 0.24$	26
Figura 14 – Exemplo de escolha do melhor limiar de rejeição para cada valor de Wr	26
Figura 15 – Regiões de decisão obtidas pela utilização de dois classificadores inde- pendentes para um problema binário. Fonte: (ROCHA NETO, 2011)	31
Figura 16 – Curvas de rejeição possíveis.	33
Figura 17 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema PVC-2C.	40
Figura 18 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema KC2.	42
Figura 19 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema HABERMAN.	44
Figura 20 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema Pima indians Diabetes.	46

Lista de tabelas

Tabela 1 – Descrição resumida dos problemas.	37
Tabela 2 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema PVC-2C.	39
Tabela 3 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema PVC-2C.	40
Tabela 4 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema KC2.	41
Tabela 5 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema KC2.	42
Tabela 6 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema HABERMAN.	43
Tabela 7 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema HABERMAN.	44
Tabela 8 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema Pima indians Diabetes.	45
Tabela 9 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema Pima indians Diabetes.	46

Lista de abreviaturas e siglas

AM	Aprendizagem de Máquina
ARC	<i>Accuracy-Rejection Curve</i>
IA	Inteligência Artificial
K-NN	<i>k Nearest Neighbors</i>
MLM	<i>Minimal Learning Machine</i>
MLM-NN	<i>MLM-Nearest Neighbors</i>
MLP	<i>Multi-Layer Perceptron</i>
RNA	Redes Neurais Artificial
wMLM	<i>Weighted Minimal Learning Machine</i>

Lista de símbolos

a	Valor de atributo
B	Matriz de regressão
C_j	j -ésima classe
D	Dimensão do espaço de entrada
\mathcal{D}_{conj}	Conjunto de distâncias entre dois pontos
$\mathcal{D}_{euclidiana}$	Distância euclidiana
\mathbf{D}_x	Contêm as distâncias entre os pontos de referência do espaço de entrada em relação aos pontos de entrada.
δ	Distância entre dois pontos de saída
$\mathbf{\Delta}_y$	Contêm as distâncias entre os pontos de referência do espaço de saída em relação aos pontos de saída.
E	Matriz de resíduos do mapeamento g
f	Mapeamento entre o espaço de entrada e saída
$F(x C_\pi)$	Função de verossimilhança dos dados da classe i
g	Mapeamento entre a matriz de distância de entrada e a matriz de distância de saída correspondente
G	Parâmetro de entrada que define o grau de polarização para uma determinada classe
Γ	Limiar de decisão de rejeição
i	Índice de instâncias de entrada ou saída
j	Índice de classes
l	Rótulo de classe desconhecida
\hat{l}	Rótulo de classe estimado
L	Conjunto de rótulos
K	Quantidade de pontos de entrada ou saída selecionados aleatoriamente

\mathbb{K}	Quantidade de pontos mais próximos usando K-NN
κ	Quantidade agrupamentos ou clusters para o K-Médias
\mathbf{m}_k	k-ésimo ponto de referência do espaço de entrada
N	Número de padrões ou observações
n	Número de execuções do classificador
$P(C_j x)$	Probabilidade a posteriori para um dado vetor de entrada x
p_i	Probabilidade a priori da i-ésima classe
q	Índice de rótulos de classe desconhecida
$\Upsilon(\phi_\pi x)$	Regra de decisão ótima
R	Conjunto de pontos de referência do espaço de entrada
\mathcal{R}	Matriz de resíduos do mapeamento f
\mathbb{R}	Conjunto dos números reais
RSS	Minimização da soma dos quadrados dos resíduos
S	Número de classes
\mathbf{W}_k	Matriz positiva simétrica
\mathbf{W}	Matriz identidade de pesos
ω	Peso sináptico do Neurônio
\mathcal{T}	Transposta de uma matriz
T	Conjunto de pontos de referência do espaço de saída
\mathbf{t}_k	k-ésimo ponto de referência do espaço de saída
T_E	Taxa de erro do classificador
T_R	Taxa de rejeição do classificador
u	Função de ativação do neurônio
W_r	Custo de rejeição
W_e	Custo de errar
W_c	Custo de acertar

\mathbf{x}_i	i-ésimo ponto de entrada
X	Conjunto de Pontos de entrada
\mathbf{y}_i	i-ésimo ponto de saída
$\hat{\mathbf{y}}$	Saída estimada do classificador
Y	Conjuntos de Pontos de saída
S	Quantidade de classes
ζ_i	i-ésimo classificador

Sumário

1	Introdução	1
1.1	Justificativa	3
1.2	Objetivo Geral	3
1.3	Objetivos Específicos	4
1.4	Organização do Trabalho	4
2	Classificadores de Padrões	5
2.1	K-Vizinhos Mais Próximos	5
2.2	K-Médias	6
2.3	Perceptron de Múltiplas Camadas	7
2.4	MLM	9
2.4.1	Formulação Básica	9
2.4.2	Regressão entre Distâncias	10
2.4.3	Estimativa da saída	12
2.4.4	MLM para Classificação	13
2.5	Máquina de Aprendizagem Mínima com Vizinhos Mais Próximos (MLM-NN)	13
2.6	Máquina de Aprendizagem Mínima Ponderada (w MLM)	14
2.7	Conclusão	15
3	Classificação com opção de rejeição	17
3.1	Definição Teórica	17
3.2	Regra de Decisão Ótima	19
3.3	Abordagens de Classificação Binária com Opção de Rejeição	22
3.4	A abordagem de um único classificador padrão	23
3.4.1	K-Médias com Opção de Rejeição	26
3.4.2	MLP com opção de rejeição	27
3.4.3	K-NN com opção de rejeição	28
3.4.4	Rejo MLM (Proposta)	29
3.4.5	Rejo MLM-NN (Proposta)	29
3.5	A abordagem de dois classificadores independentes	30
3.5.1	Rejo w MLM (Proposta)	30
3.6	Avaliação de Classificador com Opção de Rejeição	32
3.7	Conclusão	33
4	Experimentos e Resultados	35
4.1	Metodologia de Experimentos	35
4.1.1	Problemas de Classificação	35
4.1.1.1	O Problema da Coluna Vertebral	35
4.1.1.2	Predição de Defeitos de Software	36

4.1.1.3	Câncer de Mama	36
4.1.1.4	Diabetes	36
4.1.2	Validação Cruzada	37
4.1.3	Parâmetros de Classificação	37
4.1.4	Parâmetros de Classificação com Opção de Rejeição	38
4.1.5	Avaliação dos modelos	38
4.2	Resultados	39
4.2.1	Resultados PVC-2C	39
4.2.2	Resultados KC2	41
4.2.3	Resultados HABERMAN	43
4.2.4	Resultados Pima indians diabetes	45
5	Conclusão e trabalhos futuros	49
	Referências	51

1 Introdução

Ao processo de indução de uma hipótese (ou aproximação de função) a partir da experiência passada dá-se o nome de Aprendizagem de Máquina (AM), um subcampo da Inteligência Artificial (IA), dedicado ao desenvolvimento de algoritmos e técnicas que permitam o computador aprender, isto é, que permitam o computador aperfeiçoar seu desempenho em alguma tarefa (GAMA; CARVALHO, 2011). Entre as várias definições de AM na literatura, em Mitchell (1997) é definida como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.

As técnicas de AM têm sido empregadas em diversas tarefas, podendo ser organizadas de acordo com diferentes critérios (GAMA; CARVALHO, 2011). De acordo com o paradigma de aprendizado a ser adotado para lidar com uma determinada tarefa, as atividades de aprendizagem podem ser divididas em: (1) preditivas e (2) descritivas.

Em atividades preditivas o objetivo é encontrar um modelo preditivo (também denominada de função ou hipótese) por meio do conjunto de dados de treinamento capaz de associar objetos (padrões) em um número de categorias ou classes (rótulos) com base nos valores de seus atributos de entrada. É dito supervisionado porque o padrão de entrada é identificado através de um rótulo ou etiqueta. Este rótulo ou etiqueta poderá ser um valor discreto no caso de uma classificação e contínuo no caso de uma regressão. Por isso, cada objeto do conjunto de treinamento deve possuir atributos de entrada e saída.

Já o padrão em atividades descritivas é associado a um agrupamento que é aprendido normalmente com base na similaridade (ou dissimilaridade) entre os exemplos de treinamento, sem o uso de rótulos, isto é, não supervisionado e onde o número de agrupamentos muitas vezes deve ser também determinado. Na Figura 1 é apresentada uma hierarquia de aprendizado de acordo com os tipos de tarefas de aprendizado.

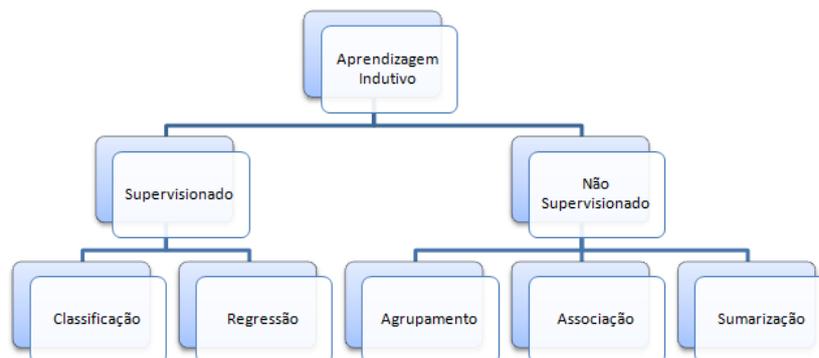


Figura 1 – Hierarquia de aprendizado (GAMA; CARVALHO, 2011)

Em problemas de aprendizado indutivo na forma supervisionada (preditivo), o

conjunto de rótulos possíveis pode ser constituído apenas por dois elementos (problema binário) ou com mais de dois elementos (problema multiclasse).

Na classificação binária de padrões complexos existem muitos padrões de classes distintas com estruturas semelhantes num espaço de características, resultando numa configuração de sobreposição entre padrões de diferentes classes. A automação das decisões nesta região de sobreposição leva, invariavelmente, para muitas previsões erradas.

Com objetivo de atingir uma maior eficácia na aprendizagem efetuada, uma nova classe pode ser definida, chamada de classe de rejeição, tal que seus padrões dispõem-se na região entre as classes com valor (+1) e (-1) em problemas do tipo binário (ROCHA NETO, 2011).

É atraente o desenvolvimento de classificadores tripartidos, com uma terceira classe de saída, isto é, a classe "rejeitado" entre os padrões da classe (+1) e (-1), porque a principal vantagem dessa estratégia é a proteção do sistema contra um eventual número de decisões erradas em situações em que se exige alta confiabilidade aplicada em situações de automatização das decisões mais difíceis. Em outras palavras, os padrões que sejam atribuídos à essa região não são classificados e são deixados para um exame minucioso sob a responsabilidade de um especialista da área ou, eventualmente, para um preditor especializado.

Entre os diversos algoritmos de aprendizagem supervisionada, a Máquina de Aprendizagem Mínima (*Minimal Learning Machine*, MLM) é um método recentemente proposto em Souza Junior et al. (2014). O funcionamento básico da MLM tem por base um mapeamento entre configurações geométricas dos pontos no espaço de entrada e saída. Com a configuração para um ponto de entrada conhecida e após a obtenção de um modelo linear entre matrizes de distância da entrada e saída, então poderá ser facilmente estimada a configuração correspondente no espaço de saída. O resultado estimado é então aproveitado para localizar o ponto de saída e, assim, prover uma estimativa para resposta ou indicação da classe.

O método de aprendizagem e suas variantes têm sido aplicados em diversos problemas de classificação e regressão conforme são demonstrados nos seguintes trabalhos (SOUZA JUNIOR et al., 2014; SOUZA JUNIOR; CORONA; BARRETO, 2014; SOUZA JUNIOR et al., 2013b; SOUZA JUNIOR et al., 2013a).

As contribuições deste trabalho são variadas, indo desde do fato de ser o primeiro a usar a Máquina de Aprendizagem Mínima e suas variantes com Opção de Rejeição aplicados em problemas de classificação binária e comparados com as técnicas MLP, K-NN e K-Médias implementadas também com capacidade de rejeição.

Portanto, a proposta deste trabalho será a aplicação do MLM e suas variantes com opção de rejeição em problemas de classificação binária.

1.1 Justificativa

A Máquina de Aprendizagem Mínima é uma técnica de aprendizagem preditiva que pode ser aplicada em problemas de regressão e classificação. O único ajuste de parâmetros é encontrar melhores pontos de referência, ou seja, um único hiperparâmetro de configuração.

O processo de ajuste dos hiperparâmetros dos classificadores tradicionais exige conhecimento tácito e pode consumir muito tempo e recursos computacionais porque é preciso encontrar os valores mais apropriados dos parâmetros livres para um determinado conjunto de dados. Na maioria dos casos, esses métodos introduzem novos parâmetros que devem ser configurados (por exemplo, o número de neurônios da camada oculta de uma Rede MLP). Além disso, ao tentar solucionar um novo problema de domínio diferente, todo o trabalho precisa ser refeito (OLIVEIRA; PRATI, 2013).

Portanto, o único hiperparâmetro do MLM o torna uma boa opção como método de aprendizagem automática em comparação à muitos métodos tradicionais que necessitam da busca de uma melhor configuração possível de vários hiperparâmetros.

Os seguintes trabalhos (SOUZA JUNIOR et al., 2014; SOUZA JUNIOR; CORONA; BARRETO, 2014; SOUZA JUNIOR et al., 2013b; SOUZA JUNIOR et al., 2013a) demonstram que o MLM tem alcançado um desempenho promissor em problemas de classificação e regressão em comparação com outros métodos clássicos de aprendizagem como Redes Neurais Artificiais (RNA) do tipo Função de Base Radial (Radial Basis Function - RBF), Perceptron Multicamadas (*Multi-Layer Perceptron* - MLP), Máquinas de Aprendizado Extremo (*Extreme Learning Machine* - ELM), Self-Organizing Map (SOM), Local Linear Map (LLM), Regional Linear Models (RLM), k Vizinhos mais próximos (K-NN) e AutoRegressivo com entrada exógena (*AutoRegressive with exogenous input* - ARX).

Além disso, aplicando o MLM com opção de rejeição será possível proteger o sistema de classificação contra erros excessivos como consequências de tomadas de decisão difícil, convertendo potenciais erros em rejeição, evitando maior confusão e delegando-os assim para a avaliação de um especialista, ou mesmo, por classificadores mais especializados.

1.2 Objetivo Geral

Em seu sentido mais amplo, o principal objetivo deste trabalho é propor a Máquina de Aprendizagem Mínima com opção de rejeição e avaliar o seu desempenho em determinados problemas de classificação binária, buscando dessa forma fornecer estratégias de aprendizagem com desempenho igual ou superior às demais técnicas convencionais.

1.3 Objetivos Específicos

Os objetivos específicos deste trabalho são apresentados a seguir.

1. Contribuir com um levantamento bibliográfico sobre o tema de interesse, enfatizando os trabalhos voltados para a classificação utilizando a Máquina de Aprendizagem Mínima;
2. Propor o método MLM e suas variantes com Opção de Rejeição;
3. Realizar um estudo reproduzível e comparativo de modelos preditivos com opção de rejeição, mais especificamente, o MLM em comparação com o MLP, K-Médias e K-NN.
4. Analisar o desempenho de classificadores com opção de rejeição quando aplicados em problemas binários

1.4 Organização do Trabalho

O restante deste trabalho é formado por cinco capítulos que descrevem os diversos conceitos relacionados à classificação com opção de rejeição e Máquina de Aprendizagem Mínima. Por fim, a metodologia proposta e resultados são descritos nos últimos capítulos.

O Capítulo 2 contém uma descrição da abordagem teórica da Aprendizagem de Máquina e dos métodos tradicionais utilizados neste trabalho. O método MLM é explanado, assim como, seus algoritmos de treinamento/teste e escolha do hiperparâmetro. Também se inclui neste capítulo as versões MLM modificadas.

Mais especificamente, o Capítulo 3 trata da estratégia de classificação com opção de rejeição. São descritos os seguintes conceitos relacionados à classificação com opção de rejeição: motivação, regra de Chow, duas abordagens de opção de rejeição em problemas de classificação binária, a forma de avaliação do desempenho dessa estratégia, o uso de custos de rejeição e a implementação dessa estratégia nos classificadores convencionais e nos métodos propostos.

O capítulo 4 é dedicado à apresentação de metodologia adotada para analisar e comparar os algoritmos de classificação citados neste trabalho. Também descreve os experimentos, tentativas, falhas e sucessos obtidos durante todo o processo de elaboração deste trabalho.

O Capítulo 5 aborda algumas considerações finais sobre esta exposição, comentando e analisando os resultados obtidos. Ressalta as contribuições que auxiliaram na elaboração desse trabalho.

2 Classificadores de Padrões

Um classificador é uma parte do software capaz de associar uma instância, cuja classe é inicialmente desconhecida, a uma classe dentro de um conjunto de finitas possibilidades. Escolher um classificador é uma tarefa difícil e essa decisão, algumas vezes, é baseada na disponibilidade do algoritmo ou no conhecimento do usuário sobre o classificador. Entretanto, a escolha correta deve-se basear em experimentações com vários classificadores, para que se possa utilizar aquele que se apresenta mais apropriado para certo tipo de problema ou aplicação.

Portanto neste capítulo, com o objetivo de facilitar a compreensão dos métodos propostos e analisar comparativamente o seu desempenho, são apresentados os classificadores de padrões utilizados neste projeto. Tomou-se como base para escrita os trabalhos de (SOUZA JUNIOR, 2014), (ROCHA NETO, 2006), (ROCHA NETO, 2011), (JAIN; DUIN; MAO, 2000), (GAMA; CARVALHO, 2011) e (HAYKIN, 1999).

O presente capítulo está estruturado em sete seções. As seções 2.1, 2.2 e 2.3 descrevem, respectivamente, os classificadores convencionais K-NN, K-Médias e MLP. A seção 2.4 descreve formalmente o funcionamento do MLM. As versões MLM-NN e w MLM são apresentadas respectivamente nas seções 2.5 e 2.6. As principais conclusões deste capítulo são relatadas na seção 2.7.

2.1 K-Vizinhos Mais Próximos

O algoritmo K-Vizinhos Mais Próximos (*k Nearest Neighbors*, K-NN) classifica os objetos com base em exemplos mais próximos no espaço de entrada. Sabendo que existe um conjunto de dados rotulados com N padrões dados por $X = \{\mathbf{x}_i\}_{i=1}^N$, com $\mathbf{x}_i \in \mathbb{R}^D$ e um padrão \mathbf{x} a ser classificado que não pertence ao conjunto X , o funcionamento do algoritmo K-NN consiste em atribuir para o padrão x o rótulo da classe mais representada nos \mathbb{K} -vizinhos mais próximos ou \mathbb{K} pontos mais próximos. Em outras palavras, é calculado um conjunto de distâncias $\mathcal{D}_{conj} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N\}$ de tal maneira que $\mathcal{D}_i = \text{distância}(\mathbf{x}, \mathbf{x}_i)$. Com base nessas distâncias são selecionadas as \mathbb{K} amostras mais próximas pertencentes ao conjunto X . Assim o padrão desconhecido é classificado pelo voto da maioria de seus vizinhos, ou seja, é classificado com a classe mais comum entre os padrões mais próximos

(ROCHA NETO, 2006). No Algoritmo 1 é descrito o funcionamento do método K-NN.

Algoritmo 1: K-Vizinhos Mais Próximos (K-NN)

- 1 Armazenar os exemplos em uma tabela;
 - 2 Seja \mathbf{x}_{new} um vetor cuja classe é desconhecida, ou seja: $Classe(\mathbf{x}_{new}) = ?$;
 - 3 Encontrar na tabela os \mathbb{K} vetores mais próximo de \mathbf{x}_{new} ;
 - 4 Seja $C_{\mathbb{K}}$ a classe a que pertence à maioria dos \mathbb{K} vetores;
 - 5 Atribuir a \mathbf{x}_{new} à classe da maioria dos \mathbb{K} vetores, ou seja $Classe(\mathbf{x}_{new}) = C_k$;
 - 6 Se a classificação for correta incluir \mathbf{x}_{new} na tabela.
-

É usual que o cálculo das distâncias \mathcal{D}_i seja feito usando a distância euclidiana

$$\mathcal{D}_{euclidiana}(\mathbf{b}, \mathbf{c}) = \sqrt{\sum_{d=1}^D (\mathbf{b}_d - \mathbf{c}_d)^2} \quad (2.1)$$

de tal maneira que D é o conjunto de características, embora outras métricas de distância possam ser utilizadas.

Este classificador tem como vantagens a simplicidade de implementação, ideal para tabelas pequenas ou médias e não requer treinamento. Além disso, constrói diretamente a regra de decisão sem estimar as densidades condicionadas às classes, sendo uma boa escolha para problemas de classificação em que padrões próximos no espaço de características possivelmente pertencem à mesma classe. Entretanto, possui como limitações o custo computacional alto para tabelas grandes e a constante \mathbb{K} é usualmente obtida por tentativa-e-erro.

2.2 K-Médias

O K-Médias (*K-Means*) é um algoritmo baseado na técnica de agrupamento (clusterização) de dados. Em outras palavras, encontra um número κ de clusters (κ é fornecido pelo usuário) que são representados por seus centróides (MACQUEEN, 1967).

Conforme é descrito pelo Algoritmo 2, o objetivo do método é separar N registros em κ agrupamentos (clusters), onde $\kappa \in \mathbb{Z}_+^*$ e $\kappa < N$. Os objetos são agrupados com base em atributos / características. Cada *cluster* é representado por um centróide. Geralmente, as coordenadas iniciais desses centróides são determinadas de forma aleatória. Cada padrão é associado a um *cluster*, também conhecido como protótipo, cujo centro está mais próximo. Essa proximidade pode ser obtida através de diversas métricas de distância, como a Euclidiana e a de Mahalanobis. Quando todos os padrões estiverem associados a algum protótipo, os κ centros são recalculados como as médias aritméticas dos padrões associados a cada agrupamento. Após encontrar os novos κ centróides, uma nova associação entre um padrão e o centro mais próximo precisa ser realizada. Portanto, essa etapa é repetida até

que as variações nos valores dos centróides sejam menores que um determinado valor.

Algoritmo 2: K-Médias

- 1 Escolha κ pontos como centróides iniciais, geralmente, de forma aleatória;
 - 2 **enquanto** *Centróides apresentam mudança* **faça**
 - 3 Forme κ grupos associando cada padrão a seu centróide mais próximo;
 - 4 Calcule o centróide de cada *cluster* usando a média aritmética;
 - 5 **fim**
-

Este algoritmo pode ser adaptado para possibilitar a sua utilização em problemas de classificações de padrões. Dessa forma, é uma técnica de aprendizagem não-supervisionada bastante utilizada devido a sua facilidade de implementação e sua ordem de complexidade $O(N)$, onde N é o número de padrões (JAIN; MURTY; FLYNN, 1999).

Como método de classificação, os padrões de treinamento são associados aos centróides mais próximos, de tal maneira que ao final deste processo estão associados κ padrões a um determinado centróide, em seguida cada centróide recebe o rótulo da classe com mais ocorrências dentre os seus κ padrões associados. Quando um exemplo não visto é apresentado, o K-Médias verifica qual o centróide mais próximo e então atribui a este padrão o rótulo deste centróide.

2.3 Perceptron de Múltiplas Camadas

Um classificador Perceptron de Múltiplas Camadas (*Multilayer Perceptron*, MLP), também conhecido por rede MLP usa o conceito de neurônio biológico para definir uma camada de entrada que recebe os vetores de atributos, e uma ou mais camadas intermediárias, compostas por neurônios somadores com função de ativação não linear e uma camada de saída, também constituída por neurônios somadores (que podem ser lineares ou não) (ROCHA NETO, 2006; CHERKASSKY; MULIER, 2007). Esse conceito de neurônio biológico segue o modelo matemático de neurônio artificial desenvolvido por McCulloch e Pitts em 1943 conforme pode ser observado na Figura 2.

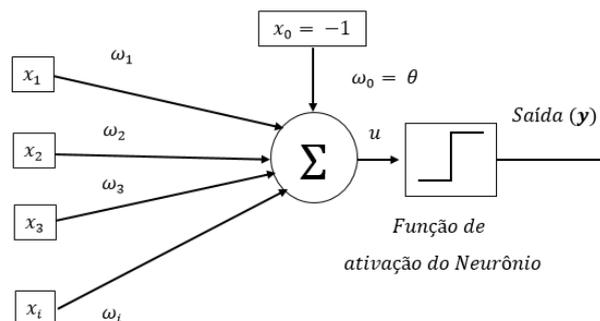


Figura 2 – Modelo de Neurônio Artificial de McCulloch e Pitts.

Na Figura 2, ω_i é o i -ésimo peso sináptico associado ao i -ésimo atributo ou característica da entrada \mathbf{x} e Θ representa o limiar (bias). A função de ativação u é calculada como $u = \omega_1 \mathbf{x}_1 + \omega_2 \mathbf{x}_2 + \dots + \omega_i \mathbf{x}_i - \Theta$.

O que habilita a rede para obter gradualmente as características mais significativas do espaço de entrada são as camadas intermediárias, principalmente, em problemas não-linearmente separáveis. Os neurônios estão interligados entre si por diferentes camadas, onde cada uma delas está ligada a um valor numérico chamado peso sináptico. Portanto, essas sinapses da rede caracterizam alto grau de conectividade dessas redes (HAYKIN, 1999). Na Figura 3 é apresentada a arquitetura geral de redes supervisionadas, como a MLP, constituída de duas camadas ocultas e um neurônio linear na saída.

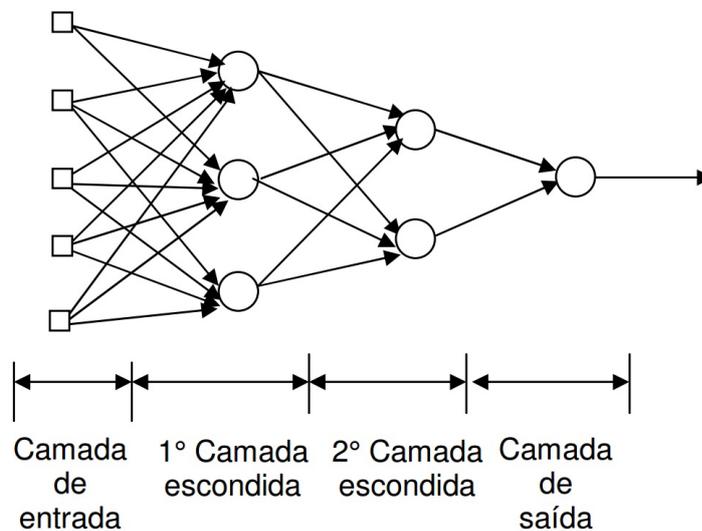


Figura 3 – Arquitetura geral de uma rede neural supervisionada (OLIVEIRA, 2001)

Equações e operações são realizadas em uma arquitetura de rede neural MLP com uma ou mais camadas ocultas de neurônios e uma aprendizagem por meio do algoritmo de retropropagação do erro (*Error Backpropagation*). Esse algoritmo é baseado na regra de correção de erro onde são ajustados os pesos dos neurônios de acordo com o erro. Dessa forma a aprendizagem da rede MLP acontece em dois sentidos: um direto e outro reverso.

O treinamento da rede MLP no sentido direto consiste no cálculo das ativações e saídas de todos os neurônios da camada oculta e de todos os neurônios da camada de saída. Portanto, a informação está se propagando no sentido direto, uma vez que o fluxo de sinais (informação) é transmitido dos neurônios de entrada para os neurônios de saída, passando apenas pelos neurônios da camada intermediária (camada escondida).

Camada de Entrada \rightarrow Camada Intermediária \rightarrow Camada de Saída

No sentido reverso o treinamento da rede MLP consiste no cálculo de gradientes locais e o ajuste dos pesos de todos os neurônios da camada de saída e da camada escondida. A informação se propaga no sentido reverso porque o fluxo de informação é transmitido

dos neurônios de saída para os neurônios da camada intermediária (camada escondida).

Camada de Saída \longrightarrow Camada Escondida

2.4 MLM

A Máquina de Aprendizagem Mínima (Minimal Learning Machine, MLM) é um algoritmo de aprendizagem indutivo supervisionado (preditivo) aplicado em problemas de classificação e regressão (SOUZA JUNIOR, 2014).

O procedimento básico no treinamento do MLM é a realização de um mapeamento entre as configurações geométricas dos pontos de entrada e seus correspondentes pontos de saídas. Nessa fase de treinamento é obtido um modelo de aprendizagem linear simples baseado em matrizes de distância da entrada e saída. Após essa fase de treinamento será possível com a configuração conhecida para um ponto de entrada estimar a configuração correspondente no espaço de saída. O resultado estimado é então aproveitado para localizar o ponto de saída e, assim, prover uma estimativa para resposta ou indicação da classe. Esse mapeamento pode ser reconstruído pela aprendizagem de um modelo de regressão linear multireposta entre distâncias.

2.4.1 Formulação Básica

Considere uma conjunto de dados de entrada rotulados com N amostras representado por $X = \{\mathbf{x}_i\}_{i=1}^N$, com $\mathbf{x}_i \in \mathbb{R}^D$, onde cada amostra há um padrão representado como um vetor $\mathbf{x} = (a_1, a_2, \dots, a_D)^T$, em que \mathcal{T} é a operação matricial de transposta e a representa um valor de atributo, com um conjunto de D características, isto é, um ponto no espaço D -dimensional. Considere também a matriz de saídas correspondente $Y = \{\mathbf{y}_i\}_{i=1}^N$, com $\mathbf{y}_i \in \mathbb{R}^S$, em que as colunas da matriz Y correspondem às S saídas e as linhas às N observações.

Adicionalmente, considera-se a existência de um mapeamento $f : X \rightarrow Y$ entre o espaço de entrada e saída, de tal forma que f seja estimado a partir dos dados com o modelo:

$$Y = f(X) + \mathcal{R}. \quad (2.2)$$

em que \mathcal{R} correspondem aos vetores de resíduo.

Dada as considerações apresentadas, pode-se descrever o funcionamento do MLM com base em duas etapas, a saber:

1. estimar o mapeamento existente entre as distâncias de entrada e saída;

2. estimar a resposta a partir da configuração dos pontos de saída.

Os dois passos são discutidos nas próximas subseções.

2.4.2 Regressão entre Distâncias

Primeiramente são selecionados pontos de entrada de referência $R = \{\mathbf{m}_k\}_{k=1}^K$ com $R \subseteq X$ com suas correspondentes saídas $T = \{\mathbf{t}_k\}_{k=1}^K$ com $T \subseteq Y$.

É definido $\mathbf{D}_x \in \mathbb{R}^{N \times K}$ de tal forma que sua k -ésima coluna $\mathbf{d}(X, \mathbf{m}_k)$ contenha as distâncias $d(\mathbf{x}_i, \mathbf{m}_k)$ entre os N pontos de entrada e o k -ésimo ponto de referência \mathbf{m}_k .

Analogamente, é definido $\mathbf{\Delta}_y \in \mathbb{R}^{N \times K}$ de forma que sua k -ésima coluna $\boldsymbol{\delta}(Y, \mathbf{t}_k)$ contém as distâncias $\delta(\mathbf{y}_i, \mathbf{t}_k)$ entre os N pontos de saída \mathbf{y}_i e a saída \mathbf{t}_k do k -ésimo ponto de referência.

O mapeamento g entre a matriz de distância de entrada \mathbf{D}_x e a matriz de distância de saída correspondente $\mathbf{\Delta}_y$ pode ser reconstruído usando o modelo de regressão multirrespostas

$$\mathbf{\Delta}_y = g(\mathbf{D}_x) + \mathbf{E}, \quad (2.3)$$

em que as colunas da matriz \mathbf{D}_x correspondem aos K vetores de entrada e as colunas da matriz $\mathbf{\Delta}_y$ correspondem aos K vetores de resposta. As colunas da matriz \mathbf{E} correspondem aos K resíduos.

Assume-se que o mapeamento g entre distâncias de entrada e saída é linear, então o modelo de regressão multireposta entre as matrizes de distância torna-se

$$\mathbf{\Delta}_y = \mathbf{D}_x \mathbf{B} + \mathbf{E}. \quad (2.4)$$

A coluna \mathbf{b}_k da matriz de regressão \mathbf{B} de dimensão $K \times K$ corresponde aos coeficientes para as K respostas. A matriz \mathbf{B} pode ser estimada a partir de dados através da minimização da soma dos quadrados dos resíduos com a função de custo:

$$\text{RSS}(\mathbf{b}_k) = (\boldsymbol{\delta}(Y, \mathbf{t}_k) - \mathbf{D}_x \mathbf{b}_k)^\top (\boldsymbol{\delta}(Y, \mathbf{t}_k) - \mathbf{D}_x \mathbf{b}_k). \quad (2.5)$$

Sob as condições normais onde o número de equações em 2.4 é maior do que o número de incógnitas, o problema em sistemas lineares é sobredeterminado e, geralmente, sem solução. Isto corresponde ao caso em que o número de pontos de referência selecionada

é menor que o número de pontos disponíveis para resolver o modelo (*i.e.*, $K < N$) sendo necessário que a solução aproximada seja fornecida pela estimativa dos mínimos quadrados:

$$\hat{\mathbf{B}} = (\mathbf{D}_x^T \mathbf{D}_x)^{-1} \mathbf{D}_x^T \mathbf{\Delta}_y. \quad (2.6)$$

Por outro lado, se em 2.4 o número de equações é igual ao número de incógnitas (*i.e.*, todos os pontos de aprendizagem são também selecionados como ponto de referência e $K = N$), o problema é unicamente determinado e, normalmente, com uma única solução $\hat{\mathbf{B}} = \mathbf{D}_x^{-1} \mathbf{\Delta}_y$. Claramente, menos interessante é o caso onde em 2.4 o número de equações é menor do que o número de incógnitas (*i.e.*, para $K > N$), sendo correspondente à situação onde, depois selecionar os pontos de referência, apenas um pequeno número de pontos de aprendizagem é usado, por isso leva a um problema subdeterminado com, normalmente, infinitas soluções.

Dada a possibilidade de $\hat{\mathbf{B}}$ ser exclusivamente resolvida ou estimada (Equação 2.6), para um ponto de treinamento $\mathbf{x}_i \in \mathbb{R}^D$ cujas distâncias de K pontos de referência $\{\mathbf{m}_k\}_{k=1}^K$ são coletadas no vetor $\mathbf{d}(\mathbf{x}, R) = [d(\mathbf{x}, \mathbf{m}_1) \dots d(\mathbf{x}, \mathbf{m}_K)]$, as distâncias correspondentes entre a saída desconhecida \mathbf{y} e as saídas conhecidas $\{\mathbf{t}_k\}_{k=1}^K$ dos pontos de referências são estimadas por

$$\hat{\delta}(\mathbf{y}, T) = \mathbf{d}(\mathbf{x}, R) \hat{\mathbf{B}}. \quad (2.7)$$

O vetor $\hat{\delta}(\mathbf{y}, T) = [\hat{\delta}(\mathbf{y}, \mathbf{t}_1) \dots \hat{\delta}(\mathbf{y}, \mathbf{t}_K)]$ fornece uma estimativa da configuração geométrica entre \mathbf{y} e o conjunto de pontos de referência T , no espaço- \mathcal{Y} .

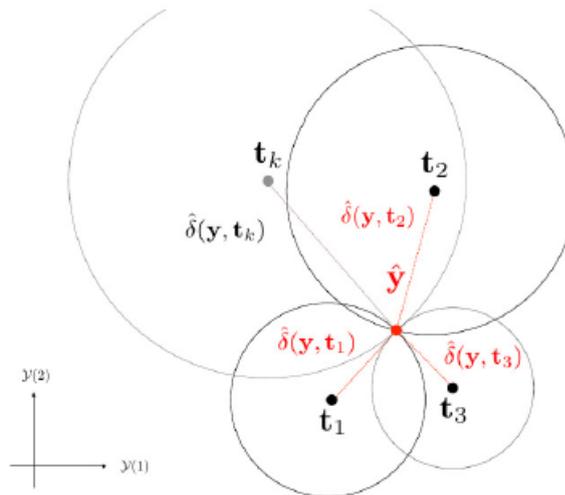


Figura 4 – Procedimento para obter estimativa.

2.4.3 Estimativa da saída

O problema de estimar a saída y , dada as saídas $\{\mathbf{t}_k\}_{k=1}^K$ para todos os pontos de referências e uma estimativa $\hat{\delta}(\mathbf{y}, T)$ das suas distâncias mútuas, pode ser entendido como um problema de multilateração para estimar sua localização em Y . O problema de localizar y é equivalente a resolver um conjunto sobredeterminado de equações não-lineares correspondentes a K hiper-esferas centradas em t_k e passando por todo y , isto é, com um raio igual a $\hat{\delta}(\mathbf{y}, \mathbf{t}_K)$:

$$(\mathbf{y} - \mathbf{t}_k)^T (\mathbf{y} - \mathbf{t}_k) = \hat{\delta}^2(\mathbf{y}, \mathbf{t}_k), \forall k = 1, \dots, K. \quad (2.8)$$

O problema na Equação 2.8 pode ser formulado como um problema de otimização onde uma estimativa $\hat{\mathbf{y}}$ pode ser obtida seguindo a minimização:

$$J(\mathbf{y}) = \sum_{k=1}^K \left((\mathbf{y} - \mathbf{t}_k)^T (\mathbf{y} - \mathbf{t}_k) - \hat{\delta}^2(\mathbf{y}, \mathbf{t}_k) \right)^2. \quad (2.9)$$

O objetivo tem um mínimo igual a 0 que pode ser alcançado se, e somente se, $\hat{\mathbf{y}}$ é a solução da Equação 2.8. Se existe, essa solução é global e única.

Devido à incerteza introduzida pelas estimativas $\hat{\delta}(\mathbf{y}, \mathbf{t}_K)$, uma solução ótima para Equação 2.9 pode ainda ser alcançada usando o método gradiente descendente ou Levenberg-Marquardt.

Os Algoritmos 3 e 4 esboçam, respectivamente, o processo de treinamento e teste.

Algoritmo 3: Algoritmo de Treinamento da MLM

Entrada: Conjunto de treinamento X , Y e K .

Saída: $\hat{\mathbf{y}}$, R e T .

- 1 Aleatoriamente selecione K pontos de referência, R , de X e suas saídas correspondentes, T , de Y ;
 - 2 Calcule D_x : As matriz de distância entre X e R ;
 - 3 Calcule D_y : As matriz de distância entre Y e T ;
 - 4 Calcule $\hat{\mathbf{B}} = (\mathbf{D}_x^T \mathbf{D}_x)^{-1} \mathbf{D}_x^T \mathbf{D}_y$;
-

Algoritmo 4: Algoritmo de Teste da MLM

Entrada: $\hat{\mathbf{B}}$, R , T e \mathbf{x} .

Saída: $\hat{\mathbf{y}}$.

- 1 Calcule $\mathbf{d}(\mathbf{x}, R)$;
 - 2 Calcule $\hat{\delta}(\mathbf{y}, T) = \mathbf{d}(\mathbf{x}, R) \hat{\mathbf{B}}$;
 - 3 Use T e $\hat{\delta}(\mathbf{y}, T)$ para encontrar uma estimativa para $\hat{\mathbf{y}}$. Isto pode ser conseguido por algum método de otimização sobre a função custo na Equação 2.9.
-

2.4.4 MLM para Classificação

Uma importante classe de problemas é classificação, na qual o objetivo é prever categorias, normalmente denotadas por saídas qualitativas (ou rótulos de classe). Novamente, tem-se disponível um conjunto de N pontos de entrada $X = \{\mathbf{x}_i\}_{i=1}^N$, com $\mathbf{x}_i \in \mathbb{R}^D$, e um conjunto de rótulos correspondentes $L = \{l_q\}_{q=1}^N$, com $l_q \in \{C_1, \dots, C_S\}$, onde C_j denota a j -ésima classe; para $S = 2$, o problema é dito ser binário, enquanto para $S > 2$ tem-se um problema multiclasse.

A Máquina de Aprendizagem Mínima pode ser especializada para problemas de classificação de uma maneira direta representando as S classes de forma vetorial através de um esquema binário de codificação (1-de- S). Nessa abordagem, uma variável qualitativa com S símbolos é representada por um vetor binário de dimensão S , dos quais somente uma componente é "ativa". Matematicamente, o conjunto de saídas $Y = \{\mathbf{y}_i\}_{i=1}^N$, com $\mathbf{y}_i \in \mathbb{R}^S$, em correspondência aos pontos de entrada X , é definida de tal forma que a j -ésima componente de \mathbf{y}_i é igual a α se $l_q = C_j$ e β caso contrário, onde α e β são inteiros com $\alpha > \beta$. Uma escolha usual é $\alpha = 1$ e $\beta = -1$.

Na etapa de classificação de uma observação de teste x com rótulo de classe desconhecido $l \in \{C_1, \dots, C_S\}$, a classe estimada \hat{l} associada com o vetor de saída estimado $\hat{\mathbf{y}}$ pelo método básico da MLM é dado por $\hat{l} = C_{s^*}$, onde

$$s^* = \arg_{s=1, \dots, S} \max \hat{y}^{(s)}. \quad (2.10)$$

As duas próximas seções apresentam uma breve descrição de duas variantes do MLM original. Ambos os métodos são usados no desenvolvimento de propostas MLMs com Opção de Rejeição.

2.5 Máquina de Aprendizagem Mínima com Vizinhos Mais Próximos (MLM-NN)

O MLM com vizinhos mais próximo (MLM-NN) consiste de um MLM modificado para a classificação onde uma busca do vizinho mais próximo é usado em vez do passo multilateração, isto é, o passo de otimização na fase de teste é evitado reduzindo, assim, sua complexidade computacional.

O MLM original foi proposto como um método geral supervisionado de aprendizagem, capaz de lidar com tarefas de classificação e regressão. Isto resultou em uma formulação geral que não aproveita as particularidades de cada tarefa.

Considerando o caso de classificação de padrões, pode-se notar que o conjunto finito e discreto de possíveis saídas podem ajudar a encontrar uma solução para a estimativa de

saída sem resolver um problema de otimização.

De fato, o passo de estimativa de saída é a parte computacionalmente crítica na MLM em comparação com os métodos convencionais de classificações. Assim, reduzir a complexidade computacional da etapa de estimativa de saída é particularmente útil no âmbito dos métodos de conjunto, uma vez que tal passo é extensivamente calculado para cada membro do conjunto.

Dessa forma, nesta proposta de trabalho uma abordagem rápida para a etapa de estimação da saída, doravante denominada MLM-NN (MLM-Nearest Neighbors), também será utilizada com opção de rejeição. Esse método é descrito formalmente em [MESQUITA, GOMES e SOUZA JUNIOR \(2015\)](#).

Nessa abordagem alternativa, a estimativa para a saída \mathbf{y} é computada com base no vizinho mais próximo (1 -NN).

Uma vez que o vetor de distâncias $\hat{\delta}(y, T)$ tenha sido estimada, pode-se utilizar a classe do ponto de referência mais próximo no espaço de saída a y para estimar a classe do padrão de teste l .

Em [MESQUITA, GOMES e SOUZA JUNIOR \(2015\)](#), é demonstrado que usando o MLM-NN para tarefas de classificação é possível obter o mesmo resultado da formulação básica MLM. Isto significa que o passo de classificação pode ser realizado com base no rótulo do ponto de referência associado com a menor estimativa de distância de saída. Na prática, são utilizadas as distâncias estimadas $\hat{\delta}(\mathbf{y}, \mathbf{t}_k)$ dadas a partir da Equação 2.7. Dessa forma, a classe estimada \hat{l} para obter um padrão de entrada \mathbf{x} é dada por $\hat{l} = l_{k^*}$, onde

$$k^* = \arg \min_{k=1, \dots, K} \{\hat{\delta}^2(\mathbf{y}, \mathbf{t}_k)\}, \quad (2.11)$$

e l_{k^*} representa o rótulo do k^* -ésimo ponto de referência.

2.6 Máquina de Aprendizagem Mínima Ponderada (w MLM)

Técnicas tradicionais de classificação supervisionada têm sido aplicadas com êxito em diversos problemas do mundo real. Geralmente, os erros são igualmente penalizados na função custo de tais métodos tradicionais. No entanto, em alguns domínios específicos, custos diferentes de erro de classificação são importantes e necessários.

Uma variante da Máquina de Aprendizagem Mínima chamada de Máquina de Aprendizagem Mínima Ponderada (Weighted Minimal Learning Machine, w MLM) foi proposta em [GOMES et al. \(2015\)](#). A ideia principal do w MLM é ponderar as instâncias do conjunto de treinamento e modificar a contribuição de cada amostra na definição do modelo final do MLM.

Este método adiciona flexibilidade ao MLM original, uma vez que é capaz de lidar com uma variedade de problemas de classificação, mantendo as vantagens MLM.

O MLM ponderado consiste numa generalização dos mínimos quadrados entre a matriz de distância de entrada \mathbf{D}_x e a matriz de distância de saída $\mathbf{\Delta}_y$ correspondente. Esta formulação traz um peso diferenciado dos resíduos na etapa de regressão entre distâncias.

O *wMLM* é obtido através da extensão do critério dos mínimos quadrados para a forma generalizada. Com a definição de uma matriz de peso \mathbf{W} , então a função de custo do MLM pode ser reescrita como:

$$RSS(\mathbf{B}) = \text{tr} \left(\mathbf{W}(\mathbf{\Delta}_y - \mathbf{D}_x \mathbf{B})^T \mathbf{W}(\mathbf{\Delta}_y - \mathbf{D}_x \mathbf{B}) \right). \quad (2.12)$$

onde \mathbf{W} é uma matriz diagonal e cada elemento W_{ii} da sua diagonal representa o peso de cada padrão de treinamento \mathbf{x}_i .

A otimização pode agora ser definida como um problema de mínimos quadrados ponderado e a matriz \mathbf{B} pode ser estimada por:

$$\hat{\mathbf{B}} = (\mathbf{D}'_x \mathbf{W} \mathbf{D}_x)^{-1} \mathbf{D}'_x \mathbf{W} \mathbf{\Delta}_y. \quad (2.13)$$

O cálculo modificado de $\hat{\mathbf{B}}$ define o *wMLM* uma vez que todos outros passos podem ser realizados de acordo com a formulação original.

2.7 Conclusão

Este capítulo apresentou formalmente o algoritmo de aprendizado supervisionado MLM inicialmente para problemas de regressão e depois em problemas de classificação. Um procedimento de estimativa de saída MLM foi descrito para reduzir a complexidade computacional, isto é, a ideia de computar a saída da rede MLM com base nas classes dos vizinhos mais próximos, tornando assim o teste da MLM mais rápido. Além dessa variante, o método *wMLM* foi apresentado formalmente, pois será utilizada na abordagem de classificação binária com opção de rejeição por meio de dois classificadores independentes.

Neste capítulo também foi apresentado algoritmos tradicionais, a saber, k-Vizinhos mais Próximos (K-NN), K-Médias e Rede MLP para problemas de duas ou mais classes.

Esses algoritmos são propostos com opção de rejeição, sendo analisado e comparado entre si o desempenho de cada método. Ressalta-se que ter conhecimento sobre o funcionamento destes algoritmos é de grande importância na compreensão final sobre este trabalho.

O capítulo seguinte é dedicado à apresentação da estratégia de classificação com opção de rejeição.

3 Classificação com opção de rejeição

Neste capítulo, com o objetivo de facilitar a compreensão do método proposto, são abordados os conceitos relacionados à Classificação com Opção de Rejeição fundamentados em trabalhos de [Chow \(1970\)](#), [Cardoso e Costa \(2007\)](#), [Rocha Neto \(2011\)](#), [Almeida \(2010\)](#), [Arlandis, Perez-Cortes e Cano \(2002\)](#).

O presente capítulo está estruturado em sete seções. A seção [3.1](#) define teoricamente a classificação com opção de rejeição citando a sua principal vantagem de proteger o sistema de apoio a decisão contra eventuais erros críticos. A descrição dessa estratégia de rejeição, isto é, rejeitar as observações menos prováveis é denotado pela regra de Chow conforme apresentado na seção [3.2](#). A seção [3.3](#) descreve os tipos de abordagens de classificação binária com opção de rejeição. As seções [3.4](#) e [3.5](#) apresentam com mais detalhes de implementação as duas abordagens aplicadas neste trabalho e a seção [3.6](#) explica a representação gráfica utilizada para medir o desempenho dessa técnica. As principais conclusões deste capítulo são apresentadas na seção [3.7](#).

3.1 Definição Teórica

Em problemas de classificação automática de padrões por meio da aprendizagem supervisionada, classificar significa atribuir um valor (classe) a um determinado padrão. Se o conjunto de classes possíveis for constituído apenas por dois elementos, então o problema é do tipo binário.

Geralmente, um classificador define em um problema binário uma saída $\mathbf{y} = +1$ ou $\mathbf{y} = -1$ com base em estimativas de probabilidades a posteriori $p(C_j|\mathbf{x})$, para um dado vetor de entrada \mathbf{x} , mesmo quando a diferença entre os valores das probabilidades a posteriori das classes é pequena. Isso resulta em uma automatização das decisões mais difíceis¹ podendo conduzir a predições erradas e, portanto, à elevação do erro de classificação ([ROCHA NETO, 2011](#)).

É comum em problema binário, rotular os vetores de entrada \mathbf{x} com valor (+1) ou (-1). Porém, é interessante em muitos ambientes que tais sistemas rejeitem casos críticos, evitando maior confusão e delegando-os assim para a avaliação de um especialista. Nesse sentido, uma nova classe pode ser definida, chamada de classe de rejeição, tal que seus padrões dispõem-se na região entre as classes com valor (+1) e (-1) ([ROCHA NETO, 2011](#)). A Figura [5](#), obtida de [Sousa, Mora e Cardoso \(2009a\)](#), exemplifica a classificação binária com opção de rejeição.

¹ Decisões em que as probabilidades a posteriori encontram-se muito próximas ao limiar.

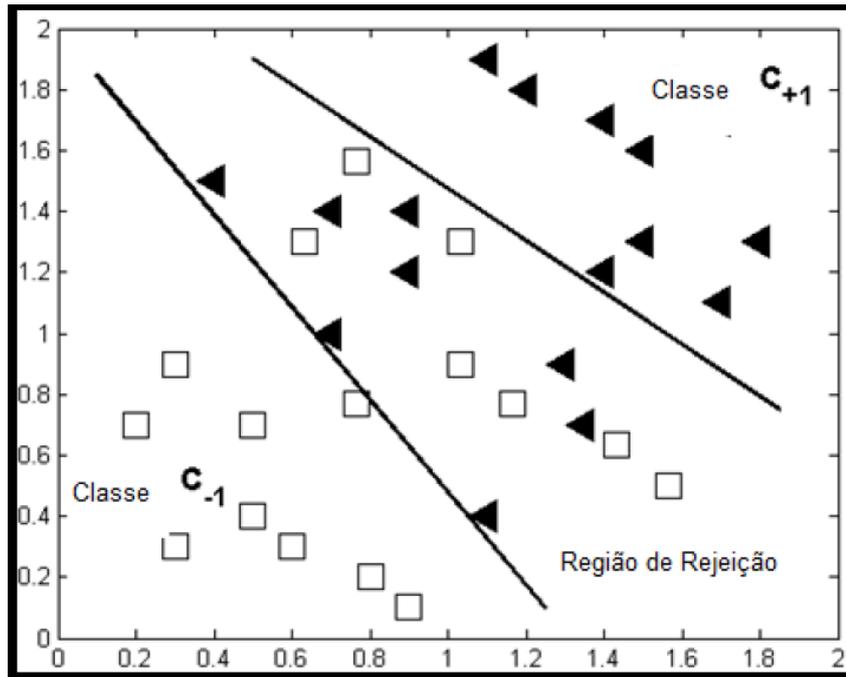


Figura 5 – Classificador Binário com opção de Rejeição. Fonte: (SOUSA; MORA; CARDOSO, 2009a)

A medida mais utilizada para avaliar o desempenho de sistemas de Reconhecimento de Padrões é a taxa de erro de um classificador. Esse valor é a razão entre o número de decisões errôneas pelo total de decisões. Outra medida importante, porém bem menos utilizada, é a taxa de rejeição, definida como a quantidade de decisões rejeitadas em relação ao número total de decisões realizadas.

Um erro significa que um padrão de uma determinada classe foi classificado como pertencente à outra classe. Quando o sistema de reconhecimento evita tomar uma decisão, é dito que o padrão foi rejeitado. Em geral, esse padrão rejeitado é posteriormente avaliado por um especialista em um exame minucioso, ou mesmo, por classificadores mais especializados. Uma descrição mais adequada do desempenho de um sistema de tomada de decisão é dada por um compromisso entre a sua taxa de erro e a sua taxa de rejeição (*trade off error-reject*)(CHOW, 1970).

Como se sabe, os erros são inevitáveis devido às incertezas e aos ruídos inerentes às tarefas de reconhecimento de padrões. A opção de rejeição serve para evitar erros excessivos como consequências de tomadas de decisão difíceis, convertendo, assim potenciais erros em rejeição. No entanto, há também a possibilidade da classificação com opção de rejeição ser utilizada em algumas decisões potencialmente corretas, ou seja, convertendo-as em rejeição. Como os custos de errar e rejeitar raramente estão na proporção um-para-um, então é preciso analisar o desempenho de predição para diferentes custos de rejeição, de tal forma, que permita escolher o melhor classificador com base no desempenho desejável ou mesmo em uma taxa de rejeição aceitável, sendo esta escolha dependente do tipo de

problema a ser abordado.

3.2 Regra de Decisão Ótima

Em um problema de classificação com opção de rejeição, uma regra de decisão é considerada ótima quando uma determinada taxa de rejeição (ou probabilidade de rejeição) minimiza a taxa de erro (ou probabilidade de erro) dado um valor de custo de rejeição Wr . A regra de decisão ótima deve recusar um determinado padrão quando o maior valor das probabilidades a posteriori é menor do que algum limiar Γ (CHOW, 1957).

Com $\mathbf{x} \in \mathbb{R}^D$ representando o padrão a ser analisado, mais explicitamente, a regra de decisão ótima $\Upsilon(\phi_\pi|\mathbf{x})$ é dada por

$$\Upsilon(\phi_\pi|\mathbf{x}) = 1 \quad (\pi \neq 0). \quad (3.1)$$

$$\Upsilon(\phi_0|\mathbf{x}) = 1 \quad (\pi = 0). \quad (3.2)$$

em que ϕ_0 representa a decisão de rejeitar o padrão x , e ϕ_π representa a decisão de aceitar o padrão \mathbf{x} e atribuí-lo à classe $\pi = 1, \dots, S$, tal que S representa o número de classes. A Equação 3.1 considera que se aceite o padrão \mathbf{x} para fins de reconhecimento e identifique-o como pertencente à π -ésima classe, se

$$p_\pi F(\mathbf{x}|C_\pi) \geq p_j F(\mathbf{x}|C_j), \forall j = 1, \dots, S \quad (3.3)$$

e

$$p_\pi F(\mathbf{x}|C_\pi) \geq (1 - \Gamma) \sum_{j=1}^S p_j F(\mathbf{x}|C_j) \quad (3.4)$$

em que C_j representa a j -ésima classe, p_j representa a probabilidade a priori da j -ésima classe, $F(\mathbf{x}|C_\pi)$ descreve a função de verossimilhança dos dados da classe π , e o limiar Γ é uma constante entre 0 e 1 ($0 \leq \Gamma \leq 1$).

A Equação 3.2, considera que se rejeite o padrão, se

$$\max_j \{p_j F(\mathbf{x}|C_j)\} < (1 - \Gamma) \sum_{j=1}^S p_j F(\mathbf{x}|C_j). \quad (3.5)$$

Definindo $m(\mathbf{x})$, tal que

$$m(\mathbf{x}) = \frac{p_j F(\mathbf{x}|C_j)}{\sum_{j=1}^S p_j F(\mathbf{x}|C_j)}, \quad (3.6)$$

então, a regra ótima de classificação pode ser simplificada e entendida como

1. aceitar o padrão \mathbf{x} , se

$$m(\mathbf{x}) \geq (1 - \Gamma), \quad (3.7)$$

ou

2. rejeitar o padrão \mathbf{x} , se

$$m(\mathbf{x}) < (1 - \Gamma), \quad (3.8)$$

Portanto, a classificação com opção de rejeição considera a rejeição do padrão \mathbf{x} sempre que o maior valor entre as probabilidades a posteriori for pequeno, ou seja, menor que $(1 - \Gamma)$ (CHOW, 1957).

O parâmetro Γ é denominado limiar de rejeição. Para qualquer valor deste limiar, a regra de decisão Υ divide o espaço de entrada em dois conjuntos (regiões) disjuntos, V_A e V_R , tal que

$$V_A(\Gamma) = \{x | m(\mathbf{x}) \geq (1 - \Gamma)\} \quad (3.9)$$

e

$$V_R(\Gamma) = \{x | m(\mathbf{x}) < (1 - \Gamma)\}. \quad (3.10)$$

A Figura 6 ilustra as regiões descritas pelos conjuntos disjuntos V_A e V_R , que são denominadas região de aceitação e região de rejeição, respectivamente. Uma região de rejeição é definida no espaço de entrada e todos os exemplos dentro desta região são rejeitados pelo classificador. Os padrões que encontram-se na região sombreada devem ser rejeitados para um dado limiar.

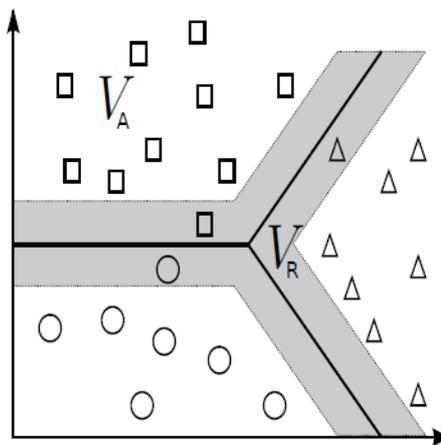


Figura 6 – Regiões de rejeição e aceitação da regra ótima de reconhecimento. Fonte: (ROCHA NETO, 2011)

Segundo Chow (1970), há duas propriedades importantes do limiar de rejeição. A primeira, refere-se à monotonicidade em Γ das taxas de erro e rejeição. A outra mostra que o parâmetro Γ é limitante superior da taxa de erro.

O desempenho de um sistema de reconhecimento é dada por uma relação de compromisso entre erro e rejeição, nas mais diversas intensidades. Uma curva típica para tal relação é apresentada na Figura 7. Pode-se observar que à medida que a rejeição aumenta a taxa de erro diminui, e vice-versa. Desde que o erro (T_E) e rejeição (T_R) de um sistema de reconhecimento ótimo são funções monotônicas do limiar de rejeição Γ , pode-se computar o compromisso entre a taxa de erro e a taxa de rejeição para um determinado limiar.

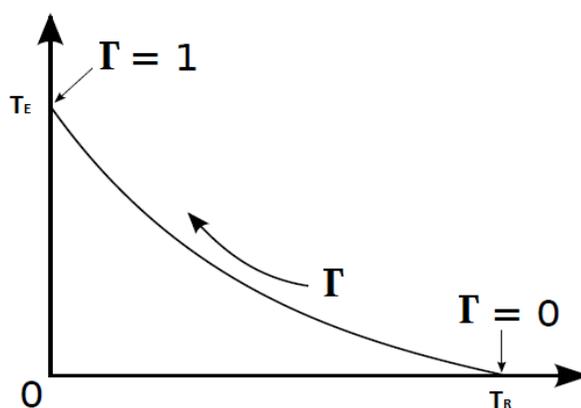


Figura 7 – Curva de compromisso entre o erro e a rejeição. Fonte: (ROCHA NETO, 2011)

Chow mostrou que a taxa de rejeição em função do limiar de rejeição descreve o desempenho do sistema de reconhecimento. Na Figura 8, obtida de (ROCHA NETO, 2011), é mostrado um caso típico deste tipo de curva. As regiões representam a taxa de erro quando nenhuma rejeição é permitida (a soma das duas áreas hachuradas), ou então a região que representa a taxa de erro quando são permitidas rejeições (região hachurada superior) (CHOW, 1970).

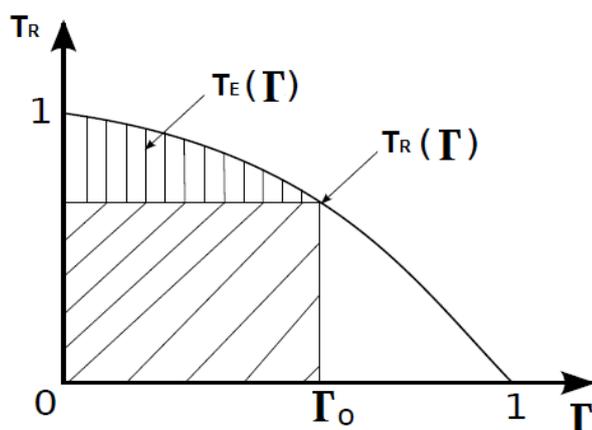


Figura 8 – Curva hipotética de compromisso entre o erro e a rejeição.

O limiar de rejeição pode então ser relacionado aos custos como segue (CHOW,

1957):

$$\Gamma = \frac{Wr - Wc}{We - Wc}. \quad (3.11)$$

em que Wr , We e Wc são os custos de rejeitar, de errar e de acertar, respectivamente. A regra de decisão ótima apresentada nas Equações 3.1 a 3.5 é também uma regra de risco mínimo se os custos associados aos erros, aos acertos e às rejeições são iguais. Usualmente, tem-se que $Wc < Wr < We$. Considerando $Wc = 0$ e $We = 1$, pode-se inferir que o custo de rejeição Wr é igual ao limiar de rejeição Γ ($Wr = \Gamma$). Assim, o risco mínimo para um determinado limiar é dado por

$$r(\Gamma) = T_E(\Gamma) + \Gamma T_R(\Gamma) = \int_0^{\Gamma_0} T_R(\Gamma) d\Gamma, \quad (3.12)$$

o que corresponde à área hachurada na Figura 8. Enquanto o risco empírico, para uma função de decisão e um determinado conjunto de treinamento, é definida como

$$r_{emp}(\alpha) = T_E(\alpha) + Wr T_R(\alpha), \quad (3.13)$$

em que α representa o vetor de parâmetros do modelo, dentre eles o limiar de rejeição.

Em um problema de classificação binária, a condição para rejeitar, descrita na Equação 3.5, pode não ser satisfeita quando $\Gamma > 0,5$, uma vez que a taxa de rejeição é sempre zero se o limiar de rejeição Γ exceder $1/2$. Esta é a razão do valor de Wr estar sempre no intervalo $[0;0,5]$.

Nas próximas seções são descritas as abordagens utilizadas neste projeto de classificadores com opção de rejeição engajados em tarefas de classificação binária. De modo geral, tais técnicas baseiam-se em 2 possibilidades: (i) otimizar o limiar de rejeição após o treinamento usual do classificador e (ii) incorporar estratégias de rejeição ao aprendizado do classificador.

3.3 Abordagens de Classificação Binária com Opção de Rejeição

Na classificação binária, o espaço das características é dividido em duas regiões pelo classificador desenvolvido, onde cada região representa uma das classes.

O classificador binário com opção de rejeição pode ser implementado através de três diferentes abordagens, a saber: (i) usando apenas um classificador; (ii) usando dois classificadores independentes; e (iii) usando um classificador com opção de rejeição incorporada na regra de aprendizado.

- **Um único classificador binário padrão:** Um padrão é rejeitado se o máximo das duas probabilidades for menor que um determinado limiar. Se o classificador não

fornecer saídas probabilísticas utiliza-se um limiar de rejeição. A região de rejeição é determinada após o treinamento do classificador, definindo valores limite adequados sobre a saída do classificador (ALMEIDA, 2010).

- **Dois classificadores independentes:** Nessa abordagem são utilizados dois classificadores denotados por ζ_1 e ζ_2 . O treinamento do classificador ζ_1 é realizado de tal forma que penalize os erros do tipo falso negativo. Por outro lado, o classificador ζ_2 é treinado com o intuito de penalizar os erros do tipo falso positivo. Desta forma, o classificador ζ_1 gera elevado valor de probabilidade para exemplos da classe C_{+1} , enquanto o classificador ζ_2 produz elevado valor para exemplos da classe C_{-1} . Um exemplo é rejeitado sempre que os classificadores discordarem (ROCHA NETO, 2011), (ALMEIDA, 2010).
- **Um único classificador com opção de rejeição incorporado:** Esta abordagem consiste em projetar classificadores que trazem embutido no processo de otimização de suas funções-custo a capacidade de classificação com opção de rejeição. Assim, não é necessário uma etapa para obtenção do limiar de rejeição ótimo, também não é necessário especializar-se uma determinada classe como ocorre na abordagem com dois classificadores independentes. Na abordagem com opção de rejeição embutida, o modelo aprende intrinsecamente a decidir entre as três saídas possíveis, C_{-1} , rejeição e C_{+1} . Alguns trabalhos relacionados a esta abordagem são apresentados em (G. FUMERA; ROLI., 2002), (A. BOUNSIAR; GRALL-MAES., 2008) e (SOUSA; MORA; CARDOSO, 2009b).

O presente trabalho adota a primeira e a segunda abordagem para a implementação de um classificador com opção de rejeição, conforme descrito em mais detalhes a seguir.

3.4 A abordagem de um único classificador padrão

A variável *limiar* é utilizada como critério para decidir a classificação ou rejeição de uma amostra. Um padrão é rejeitado se o máximo das duas probabilidades posteriori for menor que um determinado $limiar = (1 - \Gamma)$, ou seja, $\max\{p(C_{-1}|x), p(C_{+1}|x)\} < (1 - \Gamma)$. Entretanto, pode-se definir o limiar de rejeição usando uma medida de rejeição se o classificador não fornecer saídas probabilísticas. Essa medida de rejeição é calculada de acordo com o classificador utilizado.

A região de rejeição é definida após a fase de treinamento do classificador, de tal maneira que os valores de limiar adequados sobre a saída do classificador (SOUSA; MORA; CARDOSO, 2009a; ROCHA NETO, 2011). Conforme a Figura 9 mostra, o tamanho da região de rejeição é proporcional ao valor do *limiar*, isto é, quanto maior for o *limiar* então maior será a região de rejeição.

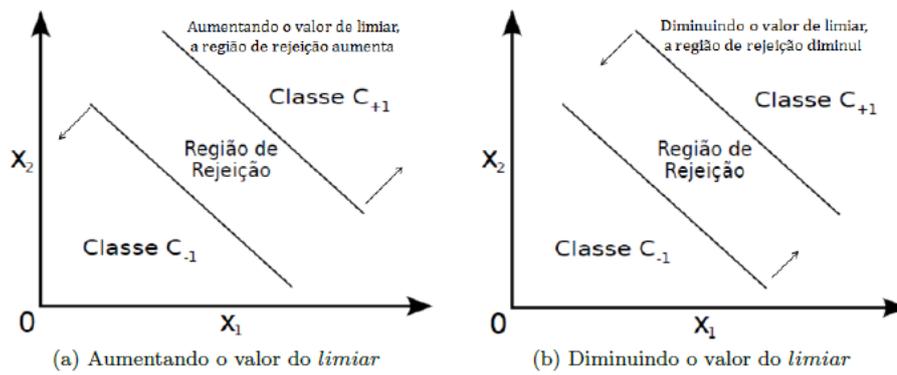


Figura 9 – Variação do tamanho da região de rejeição de acordo com o valor de *limiar*.

Na situação do *limiar* = 0 ocorre a classificação convencional, isto é, padrões não são rejeitados como a Figura 10 ilustra.

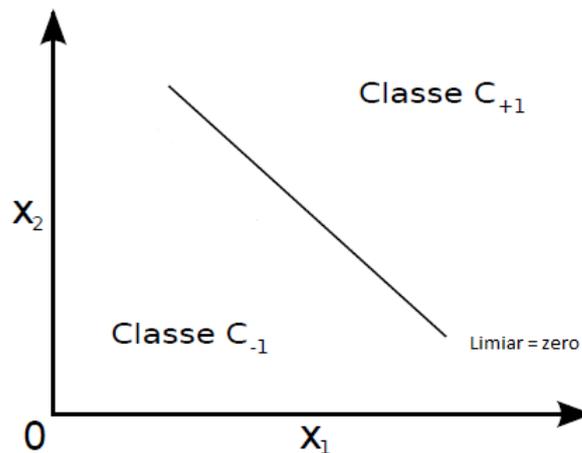


Figura 10 – Classificação convencional quando *limiar* = 0.

Apos terminar todas as n execuções do classificador vão existir n taxas de rejeição e n taxas de acerto para cada valor de *limiar*. A média de acerto e rejeição de cada *limiar* correspondente são calculadas. As Figuras 11 e 12 exemplificam esse processo de classificação com opção de rejeição para valores de *limiar* iguais a 0, 0.25 e 0.5:

A próxima etapa é escolher o *limiar* ótimo com base nos resultados obtidos, de tal forma, que permita escolher o melhor classificador com base no desempenho desejável ou mesmo em uma taxa de rejeição aceitável, sendo esta escolha dependente do tipo de problema a ser abordado.

Para isso é preciso analisar o desempenho de predição para diferentes custos de rejeição Wr buscando minimizar o risco empírico, onde T_E representa a taxa de erro e T_R sendo a taxa de rejeição (CHOW, 1957).

$$\min\{T_E(\textit{limiar}) + WrT_R(\textit{limiar})\}. \quad (3.14)$$

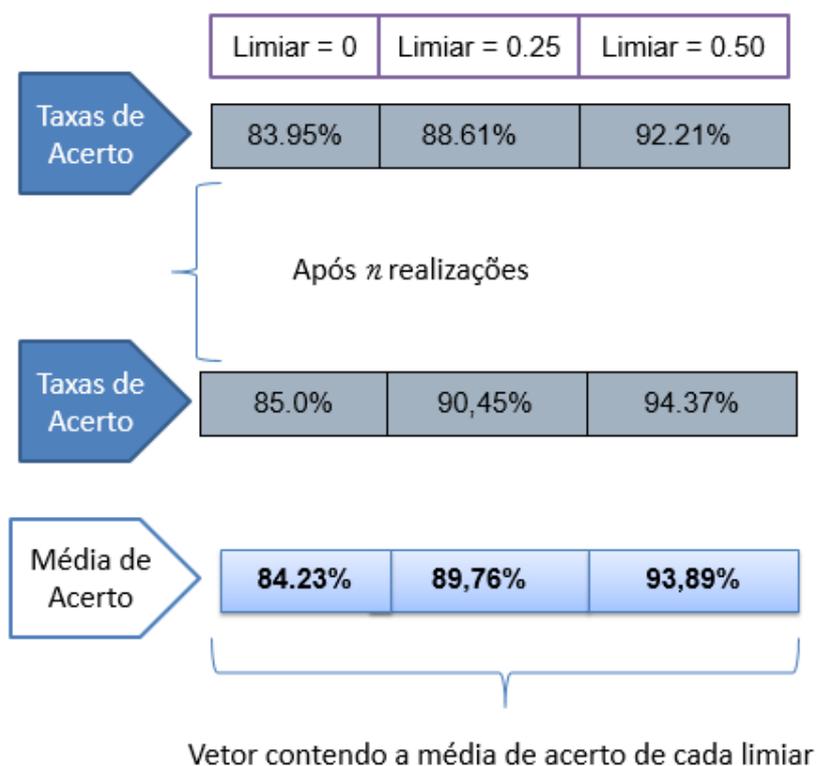


Figura 11 – Taxa de acerto para um certo *limiar* e n realizações

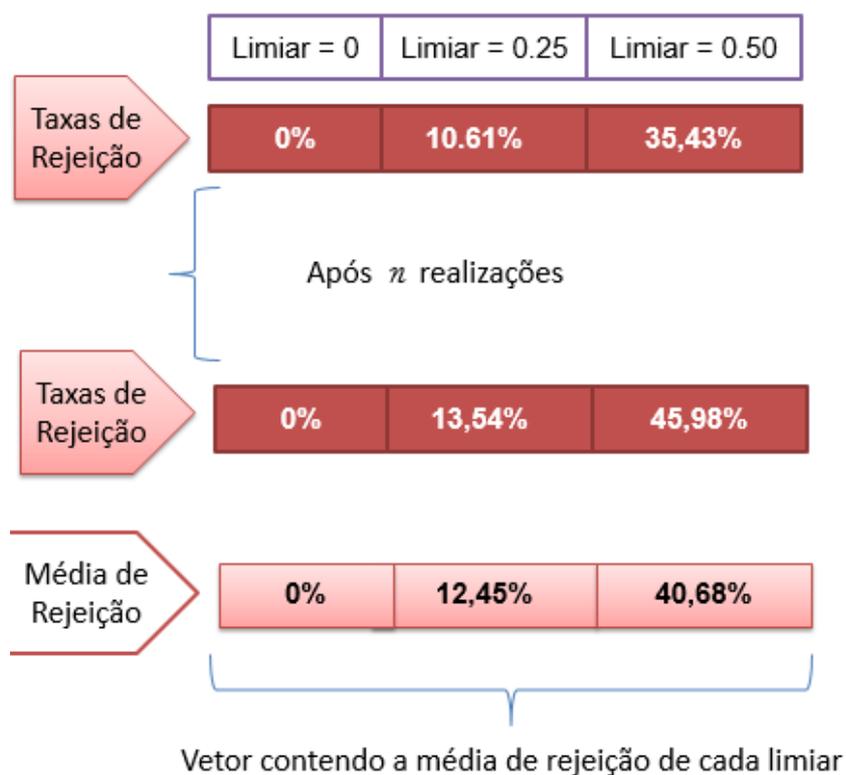


Figura 12 – Taxa de rejeição para um certo *limiar* e n realizações

O valor de Wr é inversamente proporcional a taxa de rejeição, ou seja, aumenta-se o valor de Wr quando se deseja uma menor quantidade padrões rejeitados.

Conforme a Figura 13 exemplifica para Wr igual a 0.24, o melhor resultado é quando a taxa de acerto é igual a 89,76% e a taxa de rejeição é igual a 12,45% com base nos resultados das Figuras 11 e 12.

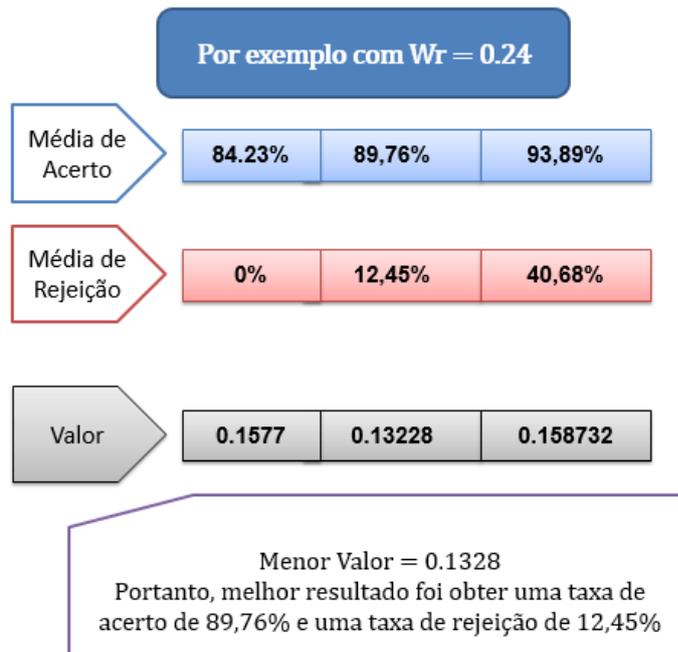


Figura 13 – Exemplo de escolha de melhor desempenho quando $Wr = 0.24$

Após obter o limiar ótimo para cada valor de Wr , é gerada uma tabela como a Figura 14 exemplifica:

Wr	Limiar ótimo	Média de acerto	Média de rejeição
0.04	0.50	93.89%	40.68%
0.24	0.25	89.76%	12.45%
0.48	0	84.23%	0%

Figura 14 – Exemplo de escolha do melhor limiar de rejeição para cada valor de Wr

3.4.1 K-Médias com Opção de Rejeição

Para classificadores com base em um quantização vetorial, tal como K-Médias, é preciso calcular uma medida de impuridade (ou heterogeneidade) da distribuição das classes dos padrões mapeados (ROCHA NETO, 2011).

A medida de impuridade é o índice de Gini que representa a desigualdade de uma distribuição (XU, 2004). Esse critério de decisão busca minimizar a impuridade de cada

protótipo, sabendo que um padrão estará com valor de impuridade máxima quando todos os padrões pertencentes ao mesmo protótipo estão igualmente distribuídos entre todas as classes e estará com valor mínimo de impuridade quando todos os padrões pertencem à mesma classe (HALMENSCHLAGER, 2002).

Seja um conjunto de dados com N padrões, cada um com uma classe C_j , então o índice de Gini (IG) é dado por:

$$IG = 1 - \sum_{j=1}^S [P(C_j|N)]^2. \quad (3.15)$$

em que $P(C_j|n)$ representa a probabilidade relativa da classe C_j em S e S é a quantidade de classes.

Conforme é descrito pelo Algoritmo 5, de acordo com índice de Gini de um aglomerado, o padrão será rejeitado ou não.

Algoritmo 5: K-Médias com opção de Rejeição

- 1 **se** $(1 - \text{índice de Gini do protótipo}) < \text{limiar}$ **então**
 - 2 | Rejeite o padrão
 - 3 **senão**
 - 4 | Classifique o padrão
 - 5 **fim**
-

3.4.2 MLP com opção de rejeição

Em classificadores como o Multilayer Perceptron (MLP) pode-se considerar a saída da rede como uma estimativa da probabilidade a posteriori $p(C_j|x)$. Neste caso, o parâmetro *limiar* varia dentro do intervalo $[0; 0.5]$ com $\Delta \text{limiar} = 0.05$ quando se utiliza a função sigmóide logística. O padrão somente será rejeitado se o valor de ativação estiver dentro desse intervalo de rejeição.

O Algoritmo 6 apresenta o algoritmo de classificação com opção de rejeição implementado no classificador MLP.

Algoritmo 6: MLP com opção de Rejeição

- 1 **se** $\text{ativação do neurônio} \in [0, 5 - \text{limiar}, 0, 5 + \text{limiar}]$ **então**
 - 2 | Rejeite o padrão
 - 3 **senão**
 - 4 | Classifique o padrão
 - 5 **fim**
-

3.4.3 K-NN com opção de rejeição

Em um classificador estatístico, uma estimativa $\hat{p}(x|C_j)$ da função de densidade de probabilidade (*probability density function*, pdf) de cada classe no espaço de características é calculada a partir de um conjunto de padrões de treinamento. No classificador original K-NN, esta função é estimada por observar o volume do espaço preenchido pelos \mathbb{K} padrões mais próximas de uma determinada classe (ARLANDIS; PEREZ-CORTES; CANO, 2002).

Na regra de votação dos k-vizinhos mais próximos, \mathbb{K} padrões são ordenados de modo crescente com base na menor distância em relação ao padrão de teste, independentemente de suas classes, e a classe mais representada é escolhida (ela é atribuído implicitamente à maior probabilidade).

Teorema de Bayes pode ser aplicada a esta estimação pdf classe-condicional para obter a estimativa de probabilidade a posteriori $\hat{P}(C_j|x)$. No entanto, em aplicações práticas, com um tamanho de amostra finito e um pequeno valor de \mathbb{K} , é muitas vezes vantajoso utilizar uma medida que leva em conta a distância entre os vizinhos mais próximos, bem como os respectivos membros da classe (ARLANDIS; PEREZ-CORTES; CANO, 2002). A estimativa é obtida por:

$$\bar{P}(C_j|x) = \frac{\sum_{i \in s_j} \frac{1}{\text{dist}(x, x_i)}}{\sum_{i=1}^k \frac{1}{\text{dist}(x, x_i)}}, \quad (3.16)$$

em que **dist** é a medida de distância usada e s_j é o conjunto de sub-índices dos protótipos de classe C_j entre os \mathbb{K} vizinhos mais próximos recuperados $x_1 \dots x_k$. Quando esta expressão é implementado, um valor pequeno ϵ adequado deve ser utilizado para **dist**(x, x_i) quando o cálculo da distância dá um resultado igual a zero.

A medida de $\hat{P}(C_c|x)$, em que c é a classe selecionada, pode ser interpretado como a confiança do classificador na sua decisão. Essa medida pode ser utilizada para rejeitar padrões que não são claramente classificados em uma classe, quando $\hat{y} < \text{limiar}$. Contudo, um padrão completamente irreconhecível ou muito ruidoso pode ser classificado com uma alta confiança quando todos ou a maioria de seus vizinhos são da mesma classe, mesmo se as distâncias envolvidas são extraordinariamente grande.

O Algoritmo 7 apresenta o algoritmo de classificação com opção de rejeição implementada no classificador K-NN.

Algoritmo 7: K-NN com opção de Rejeição

- 1 se $\hat{P}(C_i|x) < \text{limiar}$ então
 - 2 | Rejeite o padrão
 - 3 senão
 - 4 | Classifique o padrão
 - 5 fim
-

3.4.4 Rejo MLM (Proposta)

Neste projeto de classificação com opção de rejeição, é adotado a codificação 1-de-S para um conjunto de duas classes, ou seja, para um problema de classificação binária tem-se um vetor de saída de dimensão 2. Para a implementação da opção de rejeição, é necessário compor um índice capaz de medir o grau de certeza da classificação realizada, ou seja, uma medida de rejeição.

Para o método proposto foi utilizado o seguinte índice

$$Idx = \frac{\max \{\hat{y}_j\}}{\sum_{j=1}^S \hat{y}_j}. \quad (3.17)$$

em que o \hat{y}_j é o j-ésimo componente do vetor de saída \hat{y} . A função max retorna o valor do componente de \hat{y} com maior magnitude.

O índice proposto tem como base a ideia de quanto mais próximos forem os elementos do vetor \hat{y} , mais incerta será a classificação.

Para o caso da classificação binária o Idx possuirá um valor de 0.5 a 1, onde 0.5 representa máximo de incerteza e 1 representa o máximo de certeza. Dessa forma pode-se utilizar um limiar em idx para realizar a rejeição.

Portanto, para cada instância do conjunto de teste é executado o Algoritmo 8 onde há uma comparação entre o Idx calculado e o valor atual do *limiar*, sendo decidido a classificação ou rejeição da instância.

Algoritmo 8: Algoritmo para implementação do Rejo MLM

```

1 se  $Idx < \textit{limiar}$  então
2   | Rejeite o padrão
3 senão
4   | Classifique o padrão
5 fim
```

3.4.5 Rejo MLM-NN (Proposta)

Nessa versão do MLM, quando as distâncias $\hat{\delta}(y, T)$ tenham sido estimadas, antes de utilizar as classes dos K pontos de referências mais próximos no espaço de saída a y para estimar a classe do padrão de teste l , é decidido a sua classificação ou rejeição.

O critério de classificação ou rejeição é o mesmo, alterando neste caso o cálculo do Idx usado para comparação com *limiar*. Além disso, o algoritmo passa a utilizar todos os pontos de referências mais próximos no espaço de saída para o cálculo desta medida de rejeição. A inspiração para essa estratégia é a medida de rejeição utilizada pelo K-NN com

capacidade de rejeição conforme descrito pela Equação 3.16 (DALITZ; NIEDERRHEIN, 2009). Nesse contexto, o índice para a abordagem Rejo MLM-NN é dado por

$$Idx = \frac{1}{\sum_{i=1}^S \frac{1}{\min\{\text{dist}^*\}}}. \quad (3.18)$$

em que dist^* armazena todas as distâncias coletadas pelo vetor $\text{dist}(x, R)$ sem repetição de valor. A função \min retorna a menor distância do ponto de entrada do conjunto de teste em relação aos K pontos de entrada de referência.

3.5 A abordagem de dois classificadores independentes

Nesse tipo de abordagem dois classificadores denotados por ζ_1 e ζ_2 são utilizados. O treinamento do primeiro classificador ζ_1 é realizado de tal forma que penalize os erros do tipo falso negativo. Isso aumenta o nível de confiabilidade do classificador ζ_1 ao classificar um padrão como pertencente à classe C_{-1} . Por outro lado, o classificador ζ_2 é treinado com o intuito de penalizar os erros do tipo falso positivo. Assim, a certeza do padrão pertencer verdadeiramente à classe C_{+1} quando o classificador ζ_2 o classificar como pertencente à classe C_{+1} é maior. Desta forma, o classificador ζ_1 gera elevado valor de probabilidade para exemplos da classe C_{+1} , enquanto o classificador ζ_2 produz elevado valor para exemplos da classe C_{-1} (ROCHA NETO, 2011).

A regra de aceitação ou rejeição de um padrão para a estratégia baseada em dois classificadores é descrita a seguir (BOUNSIAR; GRALL; BEAUSEROY, 2007):

$$\text{SE } P(C_{+1}|x) \geq 0,5 \text{ E } P(C_{-1}|x) < 0,5 \text{ então Classe } C_{+1} \quad (3.19)$$

$$\text{SE } P(C_{+1}|x) < 0,5 \text{ E } P(C_{-1}|x) \geq 0,5 \text{ então Classe } C_{-1} \quad (3.20)$$

$$\text{SE } P(C_{+1}|x) \geq 0,5 \text{ E } P(C_{-1}|x) \geq 0,5 \text{ então rejeitar} \quad (3.21)$$

$$\text{SE } P(C_{+1}|x) < 0,5 \text{ E } P(C_{-1}|x) < 0,5 \text{ então rejeitar.} \quad (3.22)$$

Esta abordagem pode gerar uma interseção entre as superfícies de decisão das classes, resultando na ocorrência de regiões em que a confiança na tomada de decisão é bastante ruim (SOUSA; MORA; CARDOSO, 2009b). Esta situação está ilustrada na Figura 15.

3.5.1 Rejo wMLM (Proposta)

Nesse método proposto é utilizada a variante wMLM usando o algoritmo de teste do método MLM.

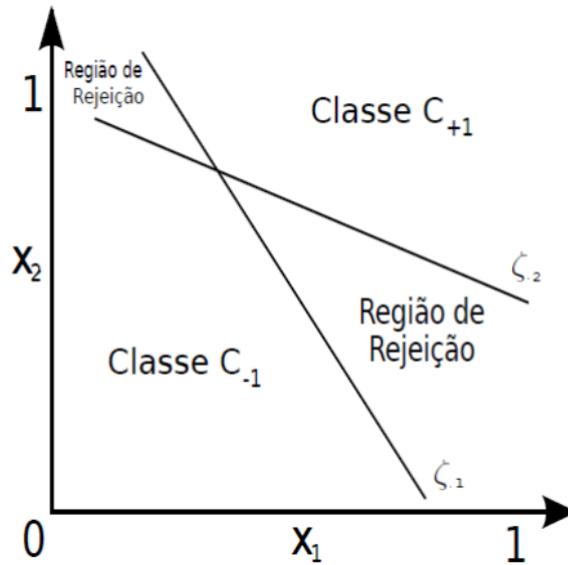


Figura 15 – Regiões de decisão obtidas pela utilização de dois classificadores independentes para um problema binário. Fonte: (ROCHA NETO, 2011)

São implementados dois classificadores ponderados denotados por ζ_1 e ζ_2 . O classificador ζ_1 gera elevado valor de confiabilidade para exemplos da classe C_{+1} , enquanto o classificador ζ_2 produz elevado valor para exemplos da classe C_{-1} . Isso é implementado através do uso da matriz diagonal W com ponderação diferente para cada classificador na fase de treinamento usando a Equação 2.13, isto é, por meio do ajuste da matriz de modo que o peso dos exemplos de classe C_{+1} são mais ponderados para a classificador ζ_1 e exemplos de classe C_{-1} são mais ponderados para o classificador ζ_2 . Um exemplo é rejeitado sempre que os classificadores discordarem.

A matriz de pesos $\mathbf{W}^{(j)}$ para j -ésimo classificador é dada por uma matriz diagonal de tal maneira que

$$w_{ii}^{(j)} = \begin{cases} G & \text{Se } \mathbf{x}_i \text{ é de classe } C_j; e \\ 1 & \text{Caso contrário;} \end{cases}$$

em que G é um parâmetro de entrada que define o grau de polarização para classe j . Neste caso, a variável de j assume apenas dois valores possíveis iguais a 1 ou 2.

Dessa forma, os elementos da diagonal principal terão valor iguais a 1 quando o elemento dessa linha não é de classe C_j e terá um valor de peso G quando é de classe C_j . Neste caso, para cada valor de polarização G são realizadas n processos de treinamento e teste do classificador.

Seja C_1 a classificação de ζ_1 e C_2 a classificação de ζ_2 , a regra de aceitação ou rejeição de um padrão para esta estratégia é descrita no Algoritmo 9:

Algoritmo 9: Algoritmo para implementação do Rejo w MLM

```

1 se  $C_1 = C_2$  então
2 | Classifique o padrão como  $C_1$ 
3 senão
4 | Rejeite o padrão
5 fim

```

3.6 Avaliação de Classificador com Opção de Rejeição

Em casos de classificação com opção de rejeição, a acurácia também depende da taxa de rejeição. Parece provável que diferentes taxas de rejeição têm efeitos diferentes sobre o desempenho de diferentes classificadores. [Nadeem, Zucker e Hanczar \(2010\)](#) descreveram um método para avaliar classificadores com opção de rejeição, apresentando os desempenhos de classificadores em curvas bidimensionais em termos de acurácia e rejeição.

Uma curva de acurácia-rejeição (*Accuracy-Rejection Curve* - ARC) é uma função que representa a taxa de acerto de um classificador em função da sua taxa de rejeição. Logo, uma curva AR terá no eixo das abscissas a Taxa de Rejeição e no eixo das ordenadas a Taxa de Acerto. Quanto maior a taxa de rejeição, maior será a taxa de acerto.

Uma curva é, por conseguinte, produzida através da representação gráfica da acurácia de um classificador contra a sua taxa de rejeição, variando de 0 a 1 (isto é, 100%). Todas as curvas têm uma acurácia de 100% para uma taxa de rejeição de 100%, e, portanto, convergem para o ponto (1,1). Elas começam a partir de um ponto (0, acc), em que acc% é a acurácia do classificador quando ele não rejeita qualquer das observações ([NADEEM; ZUCKER; HANCZAR, 2010](#)).

Curvas AR são úteis na medida em que permitem comparar graficamente as taxas de acerto de vários classificadores em função das suas taxas de rejeição. Suponha que nem todos os classificadores respondam de maneira similar em regiões diferentes de rejeição para uma dada tarefa de classificação. Alguns têm maiores taxas de acerto e menores taxas de rejeição do que outros. Com base neste pressuposto, [Nadeem, Zucker e Hanczar \(2010\)](#) identificaram três tipos diferentes de relação entre as curvas de dois classificadores comparados uns com os outros. A [Figura 16](#) ilustra o caso mais complicado que há cruzamento entre curvas AR diferentes.

Neste caso, o classificador 01 inicia com sua curva abaixo do classificador 02. Assim até antes do ponto de interseção, o classificador 02 apresenta um desempenho melhor que o classificador 01 usando a opção de rejeição, pois com uma taxa de rejeição menor consegue apresentar uma taxa de acerto maior. Após o ponto de interseção, o classificador 01 apresenta uma curva acima da curva do classificador 02 significando que apresenta um melhor desempenho nessa faixa de rejeição.

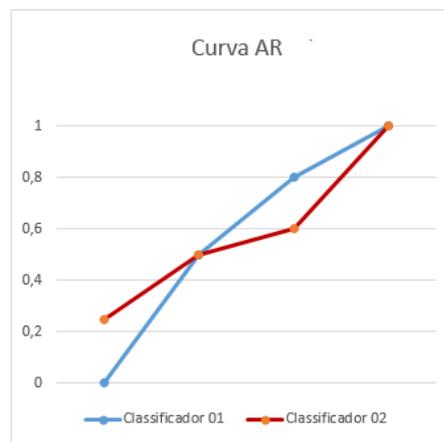


Figura 16 – Curvas de rejeição possíveis.

Esta representação gráfica, em que cada valor correspondente a uma taxa de acerto e a uma taxa de rejeição que depende do custo de rejeição W_r , é utilizada para medir o desempenho da classificação com opção de rejeição. Isto implica que diferentes pontos da curva AR correspondem a diferentes valores de W_r .

3.7 Conclusão

Este capítulo apresentou a estratégia de classificação com opção de rejeição usando a abordagem um classificador padrão. É bastante tentadora, a elaboração do problema utilizando a classificação com opção de rejeição, uma vez que protege o sistema contra um eventual número de decisões erradas em situações em que se exige alta confiabilidade.

Modelos com diferentes graus de confiança são obtidos, pois a acurácia eleva-se na medida em que a taxa de rejeição aumenta, os quais podem ser selecionados de acordo com o nível de conhecimento do especialista ou do problema de característica.

Até este capítulo foram apresentados os conceitos sobre classificação com opção de rejeição aplicada na Máquina de Aprendizagem Mínima e suas variantes. O capítulo seguinte é dedicado à apresentação da metodologia adotada e análise dos resultados obtidos.

A metodologia é avaliada através da análise comparativa de desempenho entre os métodos propostos rejeo MLM, rejeo MLM-NN e rejeo w MLM em relação aos métodos rejeo MLP, rejeo K-NN e rejeo K-Médias, sendo analisados também a taxa de rejeição e a sua eficiência através da curva AR.

4 Experimentos e Resultados

Este capítulo descreve a metodologia de avaliação e as simulações computacionais dos métodos propostos no Capítulo 3 aplicados em problemas de classificação binária.

A seção 4.1 descreve a metodologia de experimentos informando os parâmetros de classificação, implementação da opção de rejeição de modo geral, o critério de escolha dos melhores resultados e os conjuntos de dados utilizados. Os resultados dos classificadores são analisados na seção 4.2.

4.1 Metodologia de Experimentos

Nesta seção é apresentada a metodologia de experimentos dos métodos propostos aplicados em problemas de classificação binária. Cada processo de treinamento e teste do classificador foi executado 20 vezes.

4.1.1 Problemas de Classificação

O objetivo deste primeiro grupo de experimentos computacionais é avaliar o desempenho dos classificadores rejeito MLM, rejeito MLM-NN e rejeito w MLM quando aplicados ao problema de classificação. Por isso, esta seção apresenta as bases de dados utilizadas.

4.1.1.1 O Problema da Coluna Vertebral

O problema de diagnóstico de patologias da coluna vertebral está disponibilizado no repositório do site UCI Machine (UCI, 2015), onde mais detalhes sobre este problema podem ser encontrados em (ROCHA NETO, 2006).

O conjunto de dados correspondente apresenta originalmente 3 classes, a saber: normal, hérnia de disco e espondilolistese. Todavia, nesta proposta será tratado apenas o problema com 2 classes, em que se agregam os padrões pertencentes às classes hérnia de disco e espondilolistese.

Assim, o problema com 2 classes apresenta as classes normal (a mesma que no conjunto original) e anormal (com patologia). O problema com 2 classes é denominado PCV-2C (Patologias da Coluna Vertebral com 2 classes). Esse conjunto de dados possui 310 padrões, dos quais: 100 são de pacientes saudáveis e 210 de pacientes com patologia.

4.1.1.2 Predição de Defeitos de Software

Com o objetivo de incentivar a pesquisa por novos modelos preditivos de Engenharia de Software, a PROMISE (*Predictive Models In Software Engineering*) disponibiliza um repositório de bases de dados sobre problemas em Engenharia de Software (PROMISE, 2012). Nesse repositório alguns desses conjunto de dados foram criados pela NASA Metrics Data Program.

Dentre as bases disponíveis será utilizada a KC2 onde cada um dos padrões desse conjunto de dados possui 21 atributos numéricos. Ela possui 522 padrões, dos quais: 107 contêm o valor de classe igual a “Yes” e 415 contêm o valor de classe igual a “No”. O valor “Yes” é utilizado para os módulos nos quais foram encontrados defeitos e “No” para os que se apresentavam normais.

4.1.1.3 Câncer de Mama

Para este problema escolhido foi Haberman, disponibilizado no repositório do site UCI Machine (UCI, 2015).

A base Haberman tem como objetivo estudar a sobrevivência dos pacientes após a cirurgia de câncer de mama. A base possui 306 instâncias, cada uma com três atributos mais a classe a qual pertence. Os atributos são idade do paciente, ano da cirurgia e número de gânglios axilares positivos detectados. Das 306 instâncias 225 padrões refere-se à pacientes que sobreviveram 5 anos ou mais e 81 padrões aos pacientes que morreram dentro de cinco anos.

4.1.1.4 Diabetes

O último conjunto de dados utilizado é o Pima indians diabetes, disponível no repositório UCI Machine (UCI, 2015). O conjunto de dados foi filtrado para se concentrar nas pacientes mulheres com hereditariedade indiana de Pima.

Os dados incluem registros médicos como níveis de glicose e insulina, bem como fatores do estilo de vida sendo usado para prever se o indivíduo tem diabetes. Essa bases de dados possui 768 instâncias onde cada uma da instancia possui 8 atributos. Desse conjunto de dados tem-se 268 padrões que testaram positivo para diabetes e 500 padrões que testaram negativo.

A Tabela 1 sumariza os detalhes sobre cada conjunto de dados. A última coluna apresenta a taxa de desbalanceamento (IR). A IR mede o grau de desbalanceamento de cada conjunto de dados e é definido pelo número de exemplos na classe minoritária dividido pelo número de exemplos da classe majoritária.

Problema	Classe	Nº de atributos	Nº de exemplos	IR
PVC	Normal	6	100	0.4762
	Anormal		210	
KC2	Com defeito	21	107	0.2578
	Sem defeito		415	
Haberman	Sobreviveram	3	225	0.3600
	Morreram		81	
Pima indians	Positivo	8	268	0.5360
	Negativo		500	

Tabela 1 – Descrição resumida dos problemas.

4.1.2 Validação Cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados (GAMA; CARVALHO, 2011). Neste presente trabalho, foi utilizado o método *holdout*, que consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento (estimação dos parâmetros) e outro para teste (validação) (KOHAVI, 1995).

O conjunto de dados foi separado em 80% para o subconjunto de treinamento e 20% para o subconjunto de teste. A partição de treinamento é empregada para ajustar os parâmetros do modelo (classificador). O conjunto de teste é usado para estimar o erro de validação. O objetivo é obter o menor erro de generalização.

Para tanto, o classificador é treinado até atingir o erro mínimo de validação. A cada execução do classificador, as instâncias são separadas aleatoriamente sem repetição de padrões para cada subconjunto, pois é comum o conjunto de teste não incluir dados usados na fase de treinamento (DUDA; HART; STORK, 2000).

4.1.3 Parâmetros de Classificação

Os métodos de reconhecimento de padrões possuem parâmetros de classificação, tais como: número de realizações, número de vizinhos mais próximos do método K-NN, número de aglomerados do K-Médias, número de época, função de ativação, número de neurônios e camada.

O número de realizações significa a quantidade de vezes em que é realizado o procedimento de treinamento e teste. Para cada execução do algoritmo, uma nova base de treinamento e teste é obtida com seu respectivo percentual de acerto. Nesta proposta, os métodos foram executados 20 vezes.

No método K-Médias existe o parâmetro que indica o número de agrupamentos (aglomerados ou protótipos). Neste caso, K-Médias foi configurado com 3 aglomerados.

O K-NN possui o parâmetro \mathbb{K} de pontos mais próximos. O K-NN tem valor de

$\mathbb{K} = 7$.

O classificador MLP é composto por uma única camada oculta (com 30 neurônios) e um único neurônio de saída. A sigmóide logística é utilizada como função de ativação de todos os neurônios. É definido um número máximo de 100 épocas como o critério de parada. Para ajuste de cada neurônio oculto é utilizado o algoritmo retropropagação do erro (*backpropagation*).

O método MLM e suas variantes possui apenas um único hiperparâmetro que é a quantidade de pontos de referência selecionados. Neste projeto os pontos de referências escolhidos aleatoriamente representam 10% do conjunto de treinamento. Também é garantido que metade dos pontos de referência pertençam à uma classe do problema binário.

4.1.4 Parâmetros de Classificação com Opção de Rejeição

Na abordagem de um único classificador padrão, os valores atribuídos ao *limiar* variam da seguinte forma: $limiar = [0, \dots, 0.5]$. Para cada valor de *limiar* são gerados 20 taxas de acerto e rejeição correspondente ao número de execuções adotado. Por isso é calculado a média de acerto e rejeição correspondente a cada valor de limiar. Essas médias de acerto e rejeição são utilizadas para determinar o valor de *limiar* ótimo quando se utiliza um determinado custo de rejeição Wr .

Essa busca do *limiar* ótimo é realizado com base na minimização do risco empírico descrito pela Equação 3.14. As simulações foram aplicadas usando valores de custos de rejeição $Wr = [0, \dots, 0.5]$.

A metodologia na abordagem de dois classificadores independentes é semelhante à primeira abordagem. Neste caso em vez de utilizar valores de limiar são utilizados os valores atribuídos à *peso* da seguinte forma: $peso = [1, \dots, 6]$.

4.1.5 Avaliação dos modelos

O presente projeto utiliza duas métricas de avaliação, a saber: a média de acerto e a média de rejeição obtida no subconjunto de teste.

A média de acerto é a média das taxas de acertos no teste sobre o número de execuções. A taxa de acerto de cada execução é a relação entre a quantidade de instancias classificadas corretamente sobre o número de instâncias do subconjunto de teste classificadas corretamente ou não. O número de padrões rejeitados não são computados na média de acerto.

A média de rejeição é a média das taxas de rejeição sobre o número de execuções. A taxa de rejeição de cada execução é a relação entre a quantidade de instancias rejeitadas

do subconjunto de teste sobre o número total de instâncias do subconjunto teste.

A curva AR será o gráfico de desempenho gerado pelos classificadores, mantendo semelhante à forma curva em casos de bons resultados. Cada ponto da curva AR representa a média de acerto e rejeição de um determinado valor de *limiar* ótimo para um valor de Wr .

4.2 Resultados

Para avaliar o desempenho do rejeo MLM, rejeo MLM-NN e rejeo w MLM uma análise comparativa é realizada com os métodos MLP, K-Médias e K-NN com capacidade de rejeição aplicados nos problemas de classificação supracitados. As seguintes subseções descrevem os resultados obtidos em cada problema de classificação.

4.2.1 Resultados PVC-2C

Aplicando a classificação convencional, os métodos MLM, MLM-NN, K-NN, MLP e K-Médias obtiveram, respectivamente, as seguintes médias de acerto: 88.8%, 88.7%, 85.2%, 81.3% e 79.8%. Neste caso, o método MLM obteve melhor desempenho.

Conforme apresenta as Tabelas 2 e 3, o máximo de custos de rejeição Wr com valores distintos de média de acurácia (A) e de média de rejeição (R) foram: 0.02, 0.04, 0.08, 0.14, 0.22, 0.24, 0.28, 0.32, 0.38 e 0.40.

Wr	Rejo MLM		Rejo MLM-NN		Rejo w MLM	
	R	A	R	A	R	A
0.02	45.0	99.0	67.8	99.6	33.4	99.0
0.04	45.0	99.0	49.8	99.2	33.4	99.0
0.08	37.2	98.4	39.4	98.6	33.4	99.0
0.14	30.6	97.7	32.6	97.8	29.7	98.4
0.22	24.8	96.6	27.0	97.0	29.7	98.4
0.24	24.8	96.6	21.2	95.7	29.7	98.4
0.28	24.8	96.6	16.0	94.3	20.6	96.2
0.32	5.9	90.9	16.0	94.3	20.6	96.2
0.38	0.0	88.8	5.1	90.7	20.6	96.2
0.40	0.0	88.8	0.0	88.7	20.6	96.2

Tabela 2 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema PVC-2C.

W_r	Rejo MLP		Rejo K-Médias		Rejo K-NN	
	R	A	R	A	R	A
0.02	41.5	90.7	60.9	98.3	99.6	100
0.04	41.5	90.7	60.9	98.3	21.6	90.1
0.08	32.2	89.3	60.9	98.3	21.6	90.1
0.14	32.2	89.3	55.6	97.5	5.3	86.8
0.22	15.7	85.7	50.2	96.7	5.3	86.8
0.24	7.8	83.7	50.2	96.7	5.3	86.8
0.28	0.0	81.3	47.4	96.0	0.0	85.2
0.32	0.0	81.3	47.4	96.0	0.0	85.2
0.38	0.0	81.3	0.0	79.8	0.0	85.2
0.40	0.0	81.3	0.0	79.8	0.0	85.2

Tabela 3 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema PVC-2C.

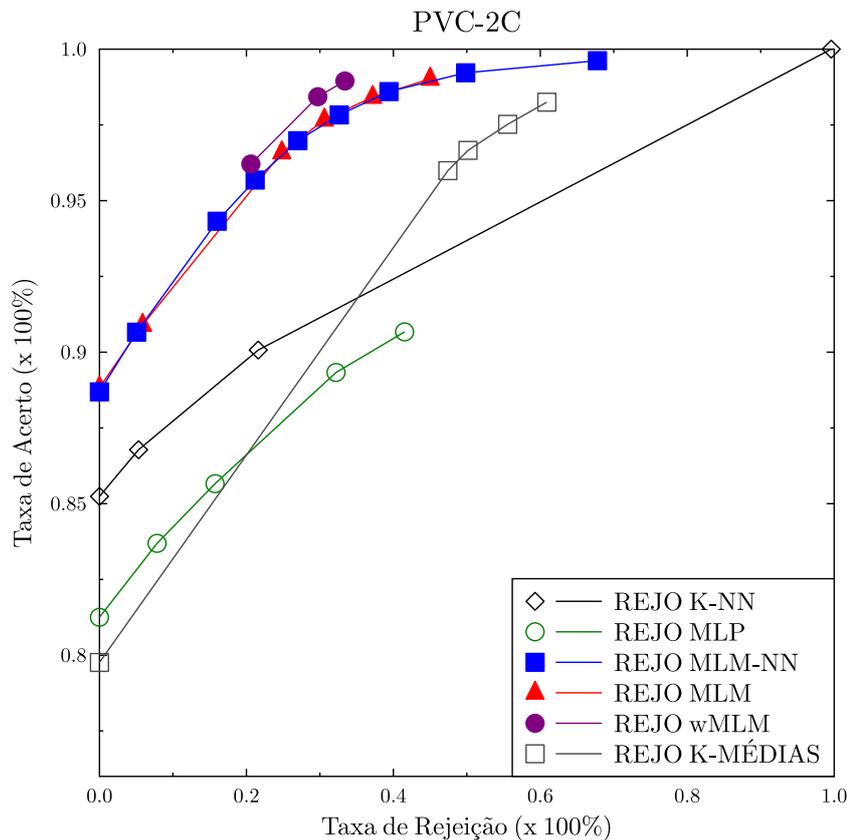


Figura 17 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema PVC-2C.

Como pode ser verificado pela Figura 17, os métodos propostos com opção de rejeição apresentaram um melhor desempenho em relação aos demais classificadores. Suas curvas AR estão acima das demais curvas indicando como melhores classificadores com opção de rejeição a ser aplicado neste tipo de problema, ou seja, significa que obtêm melhores taxas de acurácia com menor taxa de rejeição.

O rejeio w MLM apresentou o melhor desempenho na faixa de rejeição de 20% à 34% porque obteve uma melhor acurácia com uma menor média de rejeição em comparação com os demais métodos. Em seguida, o rejeio MLM-NN e rejeio MLM possui um desempenho semelhante como pode ser observado pelas suas curvas AR.

Sobre os classificadores convencionais com opção de rejeição, percebe-se que na faixa abaixo de 22% de rejeição o rejeio K-NN obteve melhor resultado em relação ao rejeio MLP e rejeio K-Médias. No intervalo de rejeição abaixo de 16% de rejeição, o rejeio K-Médias obteve pior desempenho porque teve uma alta taxa de rejeição para conseguir uma baixa taxa de acerto em comparação aos outros métodos. Após os 30% de rejeição, o rejeio K-Médias consegue apresentar melhor desempenho em relação ao rejeio MLP e rejeio K-NN.

Por ordem decrescente de desempenho, com base na Figura 17, pode-se enumerar: rejeio w MLM, rejeio MLM-NN, rejeio MLM, rejeio K-Médias, rejeio K-NN e rejeio MLP.

4.2.2 Resultados KC2

Aplicando a classificação convencional, os métodos MLM, K-Médias, K-NN, w MLM, MLP e MLM-NN obtiveram, respectivamente, as seguintes médias de acerto: 85.8%, 80.7%, 80.7%, 80.3, 78.9% e 68.7%. Neste caso, o método MLM obteve melhor desempenho.

Conforme apresenta as Tabelas 4 e 5, o máximo de custos de rejeição Wr com valores distintos de média de acurácia (A) e de média de rejeição (R) foram: 0.02, 0.04, 0.08, 0.12, 0.14, 0.16, 0.24, 0.28, 0.32, 0.38 e 0.50.

Wr	Rejeio MLM		Rejeio MLM-NN		Rejeio w MLM	
	R	A	R	A	R	A
0.02	64.9	96.8	90.0	98.0	20.0	83.6
0.04	64.9	96.8	90.0	98.0	20.0	83.6
0.12	38.2	95.4	90.0	98.0	20.0	83.6
0.16	38.2	95.4	90.0	98.0	20.0	83.6
0.20	30.7	94.2	71.3	92.9	0.0	80.3
0.24	19.1	91.7	71.3	92.9	0.0	80.3
0.28	9.3	89.2	71.3	92.9	0.0	80.3
0.32	9.3	89.2	63.5	89.5	0.0	80.3
0.40	0.0	85.8	63.5	89.5	0.0	80.3
0.48	0.0	85.8	0.0	68.7	0.0	80.3

Tabela 4 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema KC2.

Wr	Rejo MLP		Rejo K-Médias		Rejo K-NN	
	R	A	R	A	R	A
0.02	100	100	100	100	73.3	85.3
0.04	100	100	100	100	49.2	84.6
0.12	100	100	46.4	94.1	0.0	80.7
0.16	57.9	93.8	35.2	92.6	0.0	80.7
0.20	54.7	93.3	32.5	92.1	0.0	80.7
0.24	48.0	91.7	30.3	91.6	0.0	80.7
0.28	0.0	78.9	27.5	90.9	0.0	80.7
0.32	0.0	78.9	18.7	88.2	0.0	80.7
0.40	0.0	78.9	13.0	86.1	0.0	80.7
0.48	0.0	78.9	0.0	80.7	0.0	80.7

Tabela 5 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema KC2.

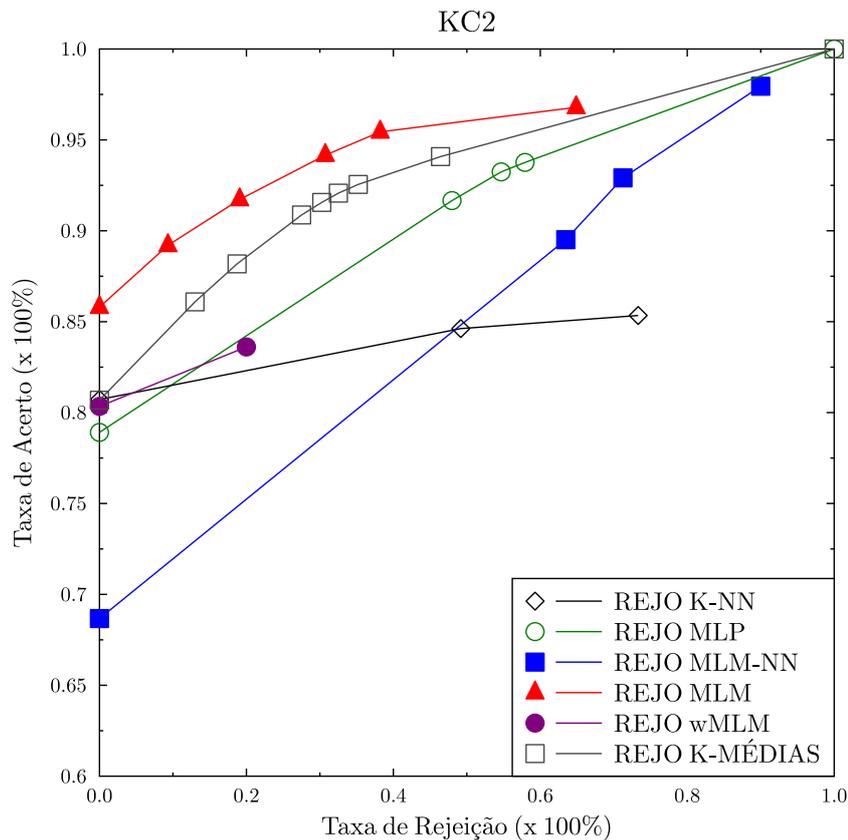


Figura 18 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema KC2.

Como pode ser observado pela Figura 18, utilizando a capacidade de rejeição dos métodos propostos apenas o reje MLM apresentou o melhor desempenho em relação a todos os classificadores. Essa conclusão é justificada pelo fato de sua curva AR estar acima das demais curvas até 65% de rejeição, ou seja, indicando ser o melhor classificador com opção de rejeição a ser aplicado neste tipo de problema.

Como pode ser visto na Tabela 4, o rejeo MLM-NN obteve o pior desempenho sem rejeição de instâncias e sua curva AR foi a mais baixa até 40% de taxa de rejeição. Após esse intervalo, o rejeo MLM-NN possui o quarto melhor resultado entre os seis classificadores analisados. O rejeo *w*MLM apresentou somente dois pontos na curva AR: (0.0%, 80.3%) e (20.0%, 83.6%). Até 20% de rejeição, a curva AR de *w*MLM foi a quarta melhor curva.

Sobre os resultados dos classificadores convencionais com opção de rejeição da Tabela 5, percebe-se que na faixa acima de 10% de rejeição é apresentado o rejeo K-Médias como o melhor método convencional com opção de rejeição. Em seguida, tem o rejeo MLP como terceiro melhor método entre todos os métodos. O rejeo K-NN apresenta pior desempenho quando a taxa de rejeição está acima de 50%.

Por ordem decrescente de desempenho, com base na Figura 18, pode-se enumerar: rejeo MLM, rejeo K-Médias, rejeo MLP, rejeo *w*MLM, rejeo MLM-NN e rejeo K-NN utilizando os custos de rejeição Wr entre 0.02 à 0.5.

4.2.3 Resultados HABERMAN

Aplicando a classificação convencional, os métodos MLM, K-Médias, K-NN, MLP e MLM-NN obtiveram, respectivamente, as seguintes médias de acerto: 79.1%, 73.4%, 73.4%, 71.9% e 64.2%. Neste caso, o método MLM obteve melhor desempenho.

Conforme apresenta as Tabelas 6 e 7, os custos de rejeição Wr com valores distintos de média de acurácia (A) e de média de rejeição (R) foram: 0.02, 0.14, 0.16, 0.20, 0.24, 0.28, 0.32, 0.36 e 0.48.

Wr	Rejo MLM		Rejo MLM-NN		Rejo <i>w</i> MLM	
	R	A	R	A	R	A
0.02	86.6	97.2	89.7	96.4	39.0	81.1
0.14	63.3	94.1	80.4	95.5	29.9	80.0
0.16	63.3	94.1	66.0	93.2	29.9	80.0
0.20	42.8	90.4	66.0	93.2	19.4	78.0
0.24	33.5	88.2	66.0	93.2	19.4	78.0
0.28	8.4	81.7	44.0	87.6	19.4	78.0
0.32	0.0	79.1	29.2	83.5	19.4	78.0
0.36	0.0	79.1	29.2	83.5	19.4	78.0
0.48	0.0	79.1	29.2	83.5	19.4	78.0

Tabela 6 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema HABERMAN.

Wr	Rejo MLP		Rejo K-Médias		Rejo K-NN	
	R	A	R	A	R	A
0.02	39.8	79.5	100	100	99,3	100
0.14	16.7	76.7	100	100	99,3	100
0.16	16.7	76.7	100	100	99,3	100
0.20	16.7	76.7	49,3	76,2	99,3	100
0.24	7.4	74.6	49,3	76,2	99,3	100
0.28	7.4	74.6	49,3	76,2	99,3	100
0.32	7.4	74.6	13,5	77,3	0,0	73,4
0.36	4.8	73.7	13,5	77,3	0,0	73,4
0.48	0,0	71.9	0,0	73,4	0,0	73,4

Tabela 7 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema HABERMAN.

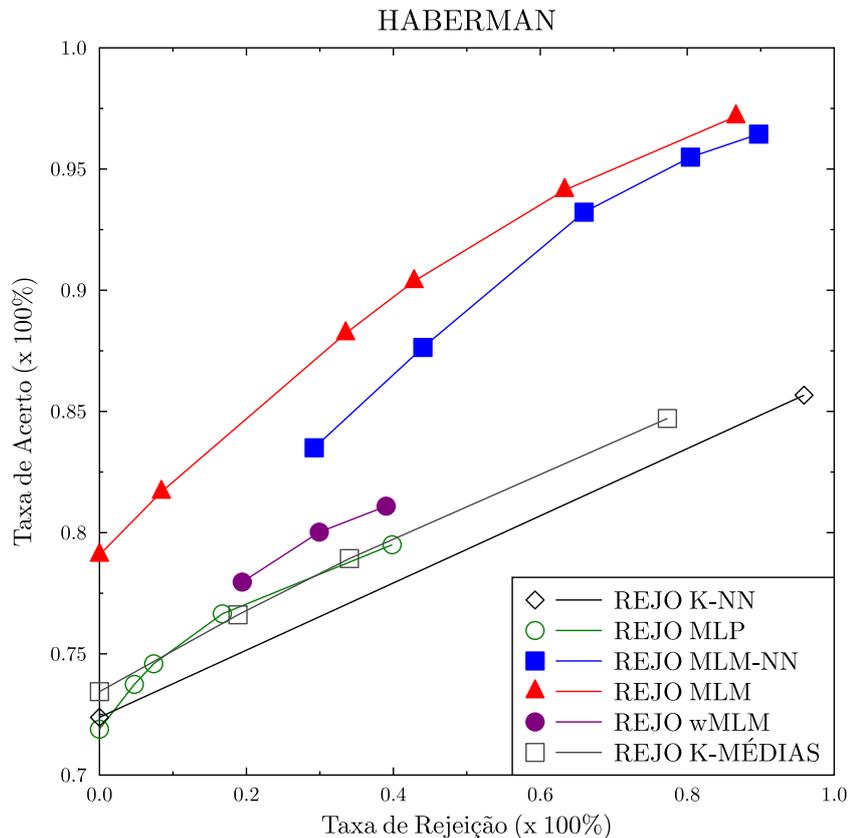


Figura 19 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema HABERMAN.

Como pode ser visto pela Figura 19, os métodos propostos com opção de rejeição apresentaram um melhor desempenho em relação aos demais classificadores, pois as curvas AR desses métodos estão acima de todas as outras curvas de métodos mais tradicionais, isto é, conseguem uma taxa de acerto melhor com uma menor taxa de rejeição. Neste caso, o reje MLM apresenta melhor resultado e depois o reje MLM-NN e wMLM.

O rejeo MLM apresenta os melhores resultados em todas as faixas de rejeição (0% à 100%), enquanto o rejeo w MLM apresenta apenas três pontos na curva entre 19% à 40% de taxa de rejeição. O rejeo MLM-NN apresenta resultados a partir de 29% de rejeição.

Em relação aos classificadores convencionais com opção de rejeição, pode-se concluir que o rejeo MLP e rejeo K-Médias tiveram um desempenho muito semelhante, pois suas curvas se interceptaram na maioria dos pontos encontrados. O rejeo K-NN obteve o pior desempenho.

Por ordem decrescente de desempenho, com base na Figura 19, pode-se enumerar: rejeo MLM, rejeo MLM-NN, rejeo w MLM, rejeo K-Médias, rejeo MLP e rejeo K-NN utilizando os custos de rejeição Wr entre 0.02 à 0.48.

4.2.4 Resultados Pima indians diabetes

Aplicando a classificação convencional, os métodos MLM, MLM-NN, K-NN, K-Médias e MLP obtiveram, respectivamente, as seguintes médias de acerto: 80.9%, 79.5%, 75.5%, 70.4% e 64.6%. Neste caso, o método MLM obteve melhor desempenho.

Conforme apresenta as Tabelas 8 e 9, os custos de rejeição Wr com valores distintos de média de acurácia (A) e de média de rejeição (R) foram: 0.04, 0.08, 0.10, 0.12, 0.14, 0.16, 0.20, 0.24, 0.28, 0.30 e 0.40.

Wr	Rejo MLM		Rejo MLM-NN		Rejo w MLM	
	R	A	R	A	R	A
0.04	84.5	96.9	83.3	95.4	56.7	95.2
0.08	71.6	95.9	83.3	95.4	56.7	95.2
0.10	71.6	95.9	70.2	94.4	56.7	95.2
0.12	61.4	94.7	59.6	93.2	56.7	95.2
0.14	61.4	94.7	59.6	93.2	56.7	95.2
0.16	51.5	93.2	41.3	90.4	56.7	95.2
0.20	34.4	90.1	41.3	90.4	52.6	94.5
0.24	26.0	88.2	33.9	88.9	37.0	91.3
0.28	9.1	83.9	25.3	86.8	37.0	91.3
0.30	9.1	83.9	0.0	79.5	37.0	91.3
0.40	0.0	80.9	0.0	79.5	37.0	91.3

Tabela 8 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método proposto aplicado no problema Pima indians Diabetes.

W_r	Rejo MLP		Rejo K-Médias		Rejo K-NN	
	R	A	R	A	R	A
0.04	24.4	72.5	97.6	93.2	100	100
0.08	24.4	72.5	97.6	93.2	100	100
0.10	24.4	72.5	97.6	93.2	100	100
0.12	24.4	72.5	97.6	93.2	100	100
0.14	24.4	72.5	84.7	91.4	100	100
0.16	24.4	72.5	71.5	89.4	100	100
0.20	24.4	72.5	71.5	89.4	0.0	75.5
0.24	24.4	72.5	71.5	89.4	0.0	75.5
0.28	24.4	72.5	0.0	70.4	0.0	75.5
0.30	9.9	68.3	0.0	70.4	0.0	75.5
0.40	0.0	64.6	0.0	70.4	0.0	75.5

Tabela 9 – Desempenho Rejeição (R) x Acurácia (A) em % de cada método convencional aplicado no problema Pima indians Diabetes.

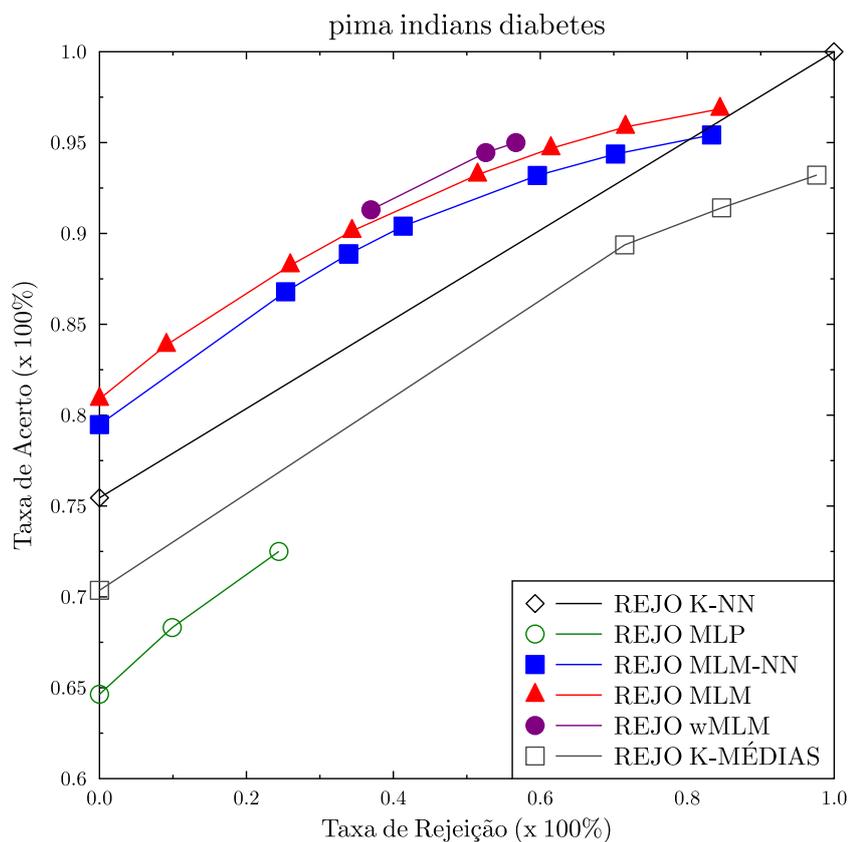


Figura 20 – Curvas de desempenho AR obtidos pelos modelos aplicados no problema Pima indians Diabetes.

Como pode ser visto pela Figura 20, os métodos propostos com opção de rejeição também apresentaram o melhor desempenho em relação aos demais classificadores, pois as curvas AR desses métodos estão acima de todas as outras curvas de métodos mais tradicionais, isto é, conseguem uma taxa de acerto melhor com uma menor taxa de rejeição.

Neste caso, o rejeio $wMLMM$ apresenta melhor resultado e depois o rejeio MLM e MLM-NN.

O rejeio $wMLM$ apresenta apenas três pontos na curva entre 37% à 56.7% de taxa de rejeição. O rejeio MLM apresenta bons resultados nas faixas de rejeição de 0% à 80%, enquanto O rejeio MLM-NN apresenta um desempenho pouco inferior.

Como pode ser visto na Tabela 9, o rejeio MLP e rejeio K-Médias não são as melhores alternativas quando se deseja rejeitar instâncias para um especialista analisar.

Por ordem decrescente de desempenho, com base na Figura 19, pode-se enumerar: rejeio $wMLM$, rejeio MLM, rejeio MLM-NN, rejeio K-NN, rejeio K-Médias e rejeio MLP utilizando os custos de rejeição Wr entre 0.04 à 0.40.

5 Conclusão e trabalhos futuros

Este trabalho apresentou um estudo na área de classificação com opção de rejeição utilizando o método MLM e suas variantes.

Nesta dissertação foram demonstrados resultados da Classificação Opção de Rejeição. Nessa estratégia, a taxa de acurácia eleva-se à medida em que a taxa de rejeição aumenta. Três classificadores foram propostos e avaliados com capacidade de rejeição: rejeição MLM, rejeição MLM-NN e rejeição w MLM. Para analisar esses resultados obtidos foram aplicados três classificadores clássicos com capacidade de rejeição: MLP, K-Médias e K-NN.

Com base nos resultados obtidos pelos classificadores aplicados pode-se concluir que, em geral, a classificação se torna mais eficiente quando são evitados erros excessivos como consequências de tomadas de decisão difíceis. Este fato é justificado, para os classificadores que implementam o algoritmo de rejeição, pelo aumento no desempenho de classificação à medida que padrões são rejeitados do conjunto de teste.

Pode-se concluir que os métodos propostos apresentaram desempenho superior em relação aos métodos rejeição K-Médias, rejeição MLP e rejeição K-NN quando aplicados aos problemas PVC-2C, HABERMAN e Pimas Indians Diabetes. Contudo, no problema KC2 apenas o rejeição MLM obteve desempenho melhor ou semelhante aos métodos clássicos, enquanto os outros métodos propostos apresentaram menor desempenho.

Nos problemas PVC-2C e Pimas Indians Diabetes, os métodos Rejeição MLM e MLM-NN tiveram um desempenho semelhante. Entretanto nos problemas KC2 e Haberman (as bases mais desbalanceadas), os métodos rejeição MLM e MLM-NN apresentaram um desempenho diferente, principalmente, no caso de KC2. O fato de KC2 ser muito desbalanceado com $IR = 0.2578$ talvez explique o desempenho do rejeição MLM-NN ter apresentado um desempenho bastante inferior ao rejeição MLM.

Portanto, este presente trabalho propõe a aplicação da técnica MLM e suas variantes na resolução de problemas de classificação com opção de rejeição. Essa estratégia de Reconhecimento de Padrões foi testada até este momento em casos como na predição de defeitos de software, no diagnóstico de patologias de coluna vertebral, câncer de mama e diabetes apresentando um desempenho semelhante ou superior às técnicas mais tradicionais da área de Aprendizagem de Máquina. A necessidade de selecionar k pontos de referência, ou seja, um único hiperparâmetro de configuração torna o MLM e suas variantes uma boa opção como método de aprendizagem automática com capacidade de rejeição.

Como futuros desdobramentos deste trabalho podem ser citados: aplicação do rejeição MLM em outros tipos de problemas.

Referências

- A. BOUNSIAR, P. B.; GRALL-MAES., E. General solution and learning method for binary classification with performance constraints. *Pattern Recognition Letters*, v. 29, p. 1455–1465, 2008. Citado na página 23.
- ALMEIDA, E. *Classificação Ordinal com Opção de Rejeição*. Dissertação (Dissertação de Mestrado) — Faculdade de Engenharia da Universidade do Porto - FEUP, 2010. Citado 2 vezes nas páginas 17 e 23.
- ARLANDIS, J.; PEREZ-CORTES, J.; CANO, J. Rejection strategies and confidence measures for a k-nn classifier in an ocr task. *IEEE*, 2002. Citado 2 vezes nas páginas 17 e 28.
- BOUNSIAR, A.; GRALL, E.; BEAUSEROY, P. A kernel based rejection method for supervised classification. *International Journal of Computational Intelligence*, v. 3, p. 312–321, 2007. Citado na página 30.
- CARDOSO, J. S.; COSTA, J. F. Learning to classify ordinal data: the data replication method. *Journal of Machine Learning Research*, v. 8, p. 1393–1429, 2007. Citado na página 17.
- CHERKASSKY, V.; MULIER, F. M. Learning from data: Concepts, theory, and methods. *Wiley-IEEE Press*, 2007. Citado na página 7.
- CHOW, C. An optimum character recognition system using decision functions. *IRE Trans. Electronic Computers, EC-B*, p. 247–254, 1957. Citado 4 vezes nas páginas 19, 20, 22 e 24.
- CHOW, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory Theory*, p. 41–46, 1970. Citado 4 vezes nas páginas 17, 18, 20 e 21.
- DALITZ, C.; NIEDERRHEIN, H. Reject options and confidence measures for knn classifiers. *Document Image Analysis with the Gamera Framework*, v. 8, p. 16–38, 2009. Citado na página 30.
- DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2nd. ed. [S.l.]: Wiley-Interscience, 2000. Citado na página 37.
- G. FUMERA; ROLI., F. Support vector machines with embedded reject option. *SVM '02: Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, p. 68–82, 2002. Citado na página 23.
- GAMA, K. F. A. C. L. J.; CARVALHO, A. C. de. *Inteligência Artificial: Uma abordagem de Aprendizado de Máquina*. 1nd. ed. [S.l.]: LTC - Livros Técnicos e Científicos, 2011. Citado 4 vezes nas páginas 15, 1, 5 e 37.
- GOMES, J. P. P. et al. A cost sensitive minimal learning machine for pattern classification. *22nd International Conference on Neural Information Processing (ICONIP2015)*, 2015. Citado na página 14.

- HALMENSCHLAGER, C. *Um algoritmo para indução de árvores e regras de decisão*. Dissertação (Dissertação de Mestrado) — Universidade Federal do Rio Grande do Sul - UFRGS, 2002. Citado na página 27.
- HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. [S.l.]: Prentice-Hall, 1999. Citado 2 vezes nas páginas 5 e 8.
- JAIN, A.; DUIN, P.; MAO, J. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. Citado na página 5.
- JAIN, A.; MURTY, M.; FLYNN, P. Data clustering: a review. *ACM Comput. Surv.*, 1999. Citado na página 7.
- KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, p. 1137–1143, 1995. Citado na página 37.
- MACQUEEN, J. Alguns métodos de classificação e análise das observações multivariada. *Proceedings de 5-th Berkeley Simpósio sobre Estatística Matemática Berkeley, University of California Press*, p. 281–297, 1967. Citado na página 6.
- MESQUITA, D. P. P.; GOMES, J. P. P.; SOUZA JUNIOR, A. H. Ensemble of minimal learning machines for pattern classification. *International Work Conference on Artificial Neural Networks, IWANN*, 2015. Citado na página 14.
- MITCHELL, T. *Machine Learning*. 1nd. ed. [S.l.]: McGraw-Hill, 1997. Citado na página 1.
- NADEEM, M.; ZUCKER, J.; HANCZAR, B. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *JMLR: Workshop and Conference Proceedings* 8, p. 65–81, 2010. Citado na página 32.
- OLIVEIRA, G. M. G.; PRATI, R. C. Ajuste de parâmetros em algoritmos de aprendizado de máquina utilizando transferência de aprendizado. *X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 2013. Citado na página 3.
- OLIVEIRA, M. P. B. *Verificação Automática do Locutor, Dependente do Texto, Utilizando Sistemas Híbridos MLP/HMM*. Dissertação (Dissertação de Mestrado) — Instituto Militar de Engenharia, 2001. Citado 2 vezes nas páginas 15 e 8.
- PROMISE. Predictive models in software engineering. 2012. Disponível em: <http://promise.site.uottawa.ca/SERepository/datasets-page.html>. Citado na página 36.
- ROCHA NETO, A. R. *SINPATCO - Sistema Inteligente para o Diagnóstico de Patologias da Coluna Vertebral*. Dissertação (Dissertação de Mestrado) — UNIVERSIDADE FEDERAL DO CEARÁ, 2006. Citado 4 vezes nas páginas 5, 6, 7 e 35.
- ROCHA NETO, A. R. *SINPATCO II: Novas Estratégias de Aprendizado de Máquina para Classificação de Patologias da Coluna Vertebral*. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO CEARÁ, 2011. Citado 10 vezes nas páginas 15, 2, 5, 17, 20, 21, 23, 26, 30 e 31.

- SOUSA, R.; MORA, B.; CARDOSO, J. An ordinal data method for the classification with reject option. *The Eighth International Conference on Machine Learning and Applications (ICMLA)*, p. 1–5, 2009. Citado 4 vezes nas páginas 15, 17, 18 e 23.
- SOUSA, R.; MORA, B.; CARDOSO, J. An ordinal data method for the classification with reject option. *Proceedings of the 2009 International Conference on Machine Learning and Applications (ICMLA '09)*, p. 746–750, 2009. Citado 2 vezes nas páginas 23 e 30.
- SOUZA JUNIOR, A. H. *Regional Models and Minimal Learning Machines for Nonlinear Dynamic System Identification*. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO CEARÁ, 2014. Citado 2 vezes nas páginas 5 e 9.
- SOUZA JUNIOR, A. H.; CORONA, F.; BARRETO, G. Minimal learning machine and local linear regression for nonlinear system identification. *Proceedings of the 20th Congresso Brasileiro de Automática*, p. 2066–2073, 2014. Citado 2 vezes nas páginas 2 e 3.
- SOUZA JUNIOR, A. H. et al. Minimal learning machine: A novel supervised distance-based method for regression and classification. *Neurocomputing*, 2014. Citado 2 vezes nas páginas 2 e 3.
- SOUZA JUNIOR, A. H. et al. Extending the minimal learning machine for pattern classification. *Proceedings 1st BRICS Countries Congress on Computational Intelligence - BRICS-CCI*, p. 236–241, 2013. Citado 2 vezes nas páginas 2 e 3.
- SOUZA JUNIOR, A. H. et al. Minimal learning machine: A new distance-based method for supervised learning. *Proceedings of the 12th International Work Conference on Artificial Neural Networks (IWANN'2013)*, p. 408–416, 2013. Citado 2 vezes nas páginas 2 e 3.
- UCI. Uci machine learning repository. 2015. Disponível em: <<https://archive.ics.uci.edu/ml/index.html>>. Citado 2 vezes nas páginas 35 e 36.
- XU, K. How has the literature on gini's index evolved in the past 80 years? *econpapers*, 2004. Citado na página 26.