UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-graduação em Engenharia Elétrica

Talles Henrique de Medeiros

Estratégias de Decisão em Aprendizado de Máquina Multi-objetivo

Belo Horizonte
2019

Talles Henrique de Medeiros

Estratégias de Decisão em Aprendizado de Máquina Multi-objetivo

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Prof. Dr. Ricardo Hiroshi Caldeira Takahashi
Coorientador: Prof. Dr. Antônio de Pádua Braga

Belo Horizonte
2019

Medeiros, Talles Henrique de.
M488e Estratégias de decisão em aprendizado de máquina multiobjetivo [recurso eletrônico] / Talles Henrique de Medeiros. - 2019.
1 recurso online (130 f. : il., color.) : pdf.

Orientador: Ricardo Hiroshi Caldeira Takahashi.
Coorientador: Antônio de Pádua Braga.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 121-130.

Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Aprendizado do computador - Teses. 3. Otimização multiobjetivo - Teses. 4. Redes neurais (Computação) - Teses. 5. Processo decisório - Teses. I. Takahashi, Ricardo Hiroshi Caldeira. II. Braga, Antônio de Pádua. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 621.3(043)

Ficha catalográfica elaborada pela bibliotecária Roseli Alves de Oliveira CRB/6 2121
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

"Estratégias de Decisão em Aprendizado de Máquina Multi-objetivo"

Talles Henrique de Medeiros

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 11 de dezembro de 2019.

Por:

Prof. Dr. Ricardo Hiroshi Caldeira Takahashi
DMAT (UFMG) - Orientador

Prof. Dr. Antônio de Pádua Braga
DELT (UFMG) - Coorientador

Prof. Dr. Frederico Gadelha Guimarães
DEE (UFMG)

Profa. Dra. Cristiane Neri Nobre
Departamento de Ciência da Computação (PUC-MG)

Prof. Dr. Marcelo Azevedo Costa
Dep. de Eng. de Produção (UFMG)

Profa. Dra. Elizabeth Fialho Wanner
Dep. Computer Science (Aston University)

Resumo

Este trabalho aborda o problema da seleção de modelos obtidos por meio do aprendizado de máquina multiobjetivo e apresenta estratégias para escolha no conjunto Pareto-ótimo. Dentro da abordagem multiobjetivo são obtidas diversas opções de soluções candidatas, caracterizadas por não-dominância entre si. Como parte da abordagem multiobjetivo, é necessário um procedimento de decisão dentre o conjunto de soluções candidatas que foram geradas. Em aprendizagem de máquina, o critério de decisão deverá retratar o dilema do equilíbrio entre os efeitos de polarização e de variância, tema central da aprendizagem de máquina e indicar uma solução que melhor represente este equilíbrio. As estratégias de decisão propostas neste trabalho foram definidas para dois dos principais problemas do aprendizado supervisionado: a classificação e a regressão. Essas estratégias caracterizam-se por serem independentes de reamostragem e da estrutura do modelo usado. A capacidade de usar novas informações para ajudar no processo de seleção do modelo garantiu avanços na abordagem do dilema entre a polarização e a variância em aprendizagem de máquina. Os resultados numéricos, por meio de problemas de aprendizagem com dados artificiais e reais, foram avaliados com outras conhecidas estratégias de decisão, como: o método da *Curva L*, o método do erro de validação e, além disso, comparados com os resultados apresentados por outros algoritmos de aprendizado como as Máquinas de Vetores Suporte (*Support Vector Machines*). Com os resultados apresentados os métodos de decisão permitiram que os algoritmos de treinamento tivessem um melhor aproveitamento do conjunto de dados original e, consequentemente, uma melhoria na capacidade de generalização. Assim, o processo de decisão em aprendizado de máquina supervisionado, sob a perspectiva da otimização multiobjetivo, trouxe um novo roteiro de seleção de modelos segundo o problema em questão, tornando o procedimento bem estruturado e determinístico.

Palavras-chave: aprendizado de máquina, otimização multiobjetivo, redes neurais artificiais, tomada de decisão.

Abstract

This work addresses the problem of selection models obtained through multi-objective machine learning and presents strategies for choosing the Pareto-optimal set. Within the multiobjective approach, several options of candidate solutions are obtained, characterized by non-dominance among themselves. As part of the multi-objective approach, a decision procedure is needed among the set of candidate solutions that have been generated. In machine learning, the decision criterion should portray the dilemma of the balance between polarization and variance effects, the central theme of machine learning, and indicate a solution that best represents this balance. The decision strategies proposed in this work were defined for two of the main problems of supervised learning: classification and regression. These strategies are characterized by being independent of resampling and the structure of the model used. The ability to use new information to help in the model selection process has ensured advances in addressing the dilemma between polarization and variance in machine learning. The numerical results, through learning problems with artificial and real data, were evaluated with other known decision strategies, such as: the *Curve L* method, the validation error method and, in addition, compared with the results presented by other learning algorithms such as the Support Vector Machines. With the results presented, the decision methods allowed the training algorithms to have a better use of the original dataset and, consequently, an improvement in the generalization capacity. Thus, the decision process in supervised machine learning, from the perspective of multiobjective optimization, brought a new model selection script according to the problem in question, making the procedure well-structured and deterministic.

Keywords: machine learning, multi-objective optimization, artificial neural networks, decision making.

Lista de Abreviaturas

ERM	Empirical Risk Minimization
RNA	Redes Neurais Artificiais
MCP	McCulloch e Pitts (modelo de RNA proposto, 1943)
MLP	Redes Neurais Multi-Camadas (*Multi-Layer Perceptron*)
MOBJ	Algoritmo Multiobjetivo (*Multi-Objective*)
MOBJ-poly	Algoritmo Multiobjetivo em Polinômios
SVM	Máquina de Vetores Suporte (*Support Vector Machine*)
RBF	Redes de Base Radial (*Radial Basis Function*)
MSE	Erro Médio Quadrado (*Mean Square Error*)
SRM	Structural Risk Minimization

l

Lista de Símbolos

x	Vetor de dados
X	Matriz do conjunto de dados de entrada
w	Classe dos dados
W	Conjunto de pesos de todas soluções de Pareto
Ω	Conjunto de soluções do problema de otimização
$\Omega(\cdot)$	Funcional estabilizador
ω	Parâmetros
$\|\|w\|\|$	Norma dos pesos sinápticos
ε	Norma dos pesos desejada
e^2	Erro quadrático do treinamento
$J(\cdot)$	Função de custo
$E_D(\cdot)$	Termo de erro padrão
$E_S(\cdot)$	Termo de regularização
λ	Parâmetro de Regularização
ζ	Número de soluções Pareto-ótimas geradas
σ^2	Variância do ruído dos dados
$f_g(\cdot)$	Função geradora dos dados
$f(\cdot)$	Função aproximada dos dados
$f_\lambda(\cdot)$	Função aproximada regularizada dos dados
ξ	Ruído branco nos dados
U	Solução Utópica
$\|\cdot\|$	Norma
α_i	Estabilizador de ordem-i de Tikhonov
$h(\cdot)$	Hipótese estatística
$L(\cdot)$	Função de Perda

Lista de Figuras

1.1 Soluções obtidas por modelos excessivamente simples e excessivamente complexos, dado um número limitado de exemplos fornecidos. 2
1.2 Exemplo de Conjunto de Soluções Pareto-ótimas e uma escolhida. 3

2.1 O Aprendizado de Máquina Supervisionado. 10
2.2 Topologia de uma Rede Multi-Layer Perceptron de uma camada oculta. . 18
2.3 Um Exemplo particular de um mapeamento das soluções no plano dos objetivos, onde a região de soluções não-dominadas está destacada em negrito. 20
2.4 Conjunto Pareto-ótimo. 25

3.1 Representação de um Problema Direto e Inverso. 31
3.2 Ideia Básica da Teoria da Regularização 32
3.3 *Curvas-L* para problemas contínuos e discretos. 36
3.4 Uma *Curva L* típica e um gráfico da curvatura k em função do parâmetro de regularização λ. 37
3.5 Curva de Pareto e a Curva de Validação. 38
3.6 Deslocamento da região de mínimo erro na curva de validação em problemas com ruído não correlacionado. 39
3.7 Deslocamento da região de mínimo erro na curva de validação em problemas com ruído correlacionado. 40
3.8 A *Curva L* (Pareto em escala logarítmica) e a melhor aproximação obtida por uma estratégia de decisão. 42
3.9 A *Curva L* (Pareto em escala logaritmica) e a melhor aproximação obtida por uma estratégia de decisão. 42
3.10 Distribuição das Classes 43
3.11 O Conjunto Pareto e a *Curva L* (Logarítmica) - Problema Gaussianas. . . 44

3.12 Curva de Decisão da melhor solução multiobjetivo, localizada na *Curva L* 44
3.13 Problema de Classificação em espiral 45
3.14 Superfície gerada pelo método multiobjetivo 46
3.15 Conjunto Pareto e a *Curva L* (Logarítmica) - Problema Espiral. 46
3.16 Problema de Regressão Não-linear 47
3.17 Curva de Pareto (Linear) e a *Curva L* (Logarítmica) - Problema Regressão Não-Linear. 47

4.1 Classificações de Máquinas Alternativas: Exata (super-ajustada) e regular (ajustada). 53
4.2 A Distribuição Binomial para $n = 40$ e $p = 0.3$. 56

5.1 Curvas de Separação e *Curva L* gerada com dados do treinamento além da escolha com decisor probabilístico com $p = 0.03$. 59
5.2 Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.03$. 60
5.3 Curvas de Separação e *Curva L* gerada com dados do treinamento além da escolha com decisor probabilístico com $p = 0.05$. 60
5.4 Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.05$. 61
5.5 Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.1$. 61
5.6 Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.1$. 62
5.7 Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.2$. 62
5.8 Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.2$. 63
5.9 Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.5$. 63
5.10 Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.5$. 64
5.11 Curvas de separação: Decisão probabilística com 10 e 1000 exemplos X SVM treinada com 1000 exemplos. 66
5.12 Curvas de separação: Decisão por Validação com 10 e 1000 exemplos X SVM treinadacom 1000 exemplos. 67

5.13 Redes mais prováveis de representar as classes +1 (+) e -1 (*), respectivamente. 68
5.14 Conjunto Pareto, em escala logarítmica, dos modelos obtidos com conjuntos de 10 e 1000 exemplos e decisão probabilística. 69
5.15 Conjunto Pareto, em escala logarítmica, dos modelos obtidos com conjuntos de 10 e 1000 exemplos e decisão por Validação. 70
5.16 Curvas de separação: Decisão por Validação apresenta *overfitting* e p = 0.005. 71
5.17 Curvas de separação: Decisão por Validação e por Hipótese, com p = 0.01. 71
5.18 Curvas de separação: Decisão por Validação e por Hipótese, com p = 0.02. 72
5.19 Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.05$. 75
5.20 Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.05$. 75
5.21 Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.30$. 76
5.22 Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.30$. 77
5.23 Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.10$. 77
5.24 Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.10$. 78
5.25 Aproximações das Distribuições Binomiais indicando as probabilidades de cada rede para a classe dos cardíacos e não-cardíacos. 82
5.26 Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.05) e por erro de validação (70% dos exemplos de treinamento) . . . 83
5.27 A *Curva L* de redes obtidas com estratégia de decisão probabilísticas (p=0.05) e por erro de validação (70% dos exemplos de treinamento) . . . 83
5.28 Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.01) e por erro validação (70% dos exemplos de treinamento) 87
5.29 Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.4) e por erro de validação (70% dos exemplos de treinamento) 88
5.30 Desempenho das redes MOBJ no conjunto de dados de teste. 88

7.1 Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.1$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso. 96
7.2 Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.1$. Este conjunto de teste é composto por 500 dados. 97
7.3 Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.2$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso. 97
7.4 Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.2$. Este conjunto de teste é composto por 500 dados. 98
7.5 Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.3$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso. 98
7.6 Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.3$. Este conjunto de teste é composto por 500 dados. 99
7.7 Função geradora $\frac{(x-2)(2x+1)}{(1+x^2)}$ com o ruído correlacionado ξ_{corr}, exibidos separadamente, e as amostras da função com ruído. 100
7.8 O ruído correlacionado gerado e a função de autocorrelação deste ruído. . 101
7.9 O ruído branco e a função de autocorrelação deste ruído. 101
7.10 A *Curva L* com a decisão por mínima correlação e aproximação correspondente à solução escolhida por mínima correlação com dados de ruídos correlacionados. 102

7.11 A Curva com os erros de teste de cada solução Pareto-ótima em relação à dados que não foram corrompidos com ruído. Aqui a solução com mínimo erro indica a localização da melhor aproximação da rede MLP em relação ao conjunto de dados disponível neste caso. 103
7.12 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 15 dados de treinamento. 104
7.13 Aproximações obtidas com o treinamento usando 15 dados e a curva do erro de teste x norma utilizando 300 dados de teste. 105
7.14 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento. 105
7.15 Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste. 106
7.16 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento. 106
7.17 Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste. 107
7.18 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 15 dados de treinamento. 108
7.19 Aproximações obtidas com o treinamento usando 15 dados e a curva do erro de teste x norma utilizando 200 dados de teste. 109
7.20 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento. 109
7.21 Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste. 110
7.22 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento. 110
7.23 Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste. 111
7.24 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento. 112
7.25 Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste. 112
7.26 Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento. 113
7.27 Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste. 114

Lista de Tabelas

5.1 Matriz de Confusão do classificador SVM. 85
5.2 Matriz de Confusão do classificador obtido por decisão probabilística. . . . 86
5.3 Matriz de Confusão do classificador obtido por erro nos dados de validação. 86
5.4 Sensibilidade do decisor probabilístico à variações do valor p 89

7.1 Erro (MSE) no conjunto de teste de 2000 dados - Problema $f_2(x)$ 115
7.2 Erro (MSE) no conjunto de teste de 2000 dados - Problema $f_3(x)$ 115

Sumário

Lista de Abreviaturas **7**

Lista de Símbolos **8**

Lista de Figuras **9**

Lista de Tabelas **14**

1 Introdução **1**
1.1 A Teoria do Aprendizado de Máquina 1
1.2 O Controle de Complexidade 3
1.3 O Problema 4
1.4 Objetivo Principal 4
1.4.1 Objetivos Específicos 5
1.5 Motivação 5
1.6 Justificativa 6
1.7 Contribuições 6
1.8 Organização do Texto 7

2 Decisão Multiobjetivo **9**
2.1 Aprendizado de Máquina e a Generalização 9
2.1.1 Princípios do Aprendizado de Máquina 9
2.1.2 Capacidade de Generalização 11
2.1.3 Seleção de Modelos 12
2.1.4 Redes Neurais MLP 17
2.2 A Otimização Multiobjetivo 18
2.2.1 O Conceito Pareto - ótimo 19
2.3 O Aprendizado de Máquina Multiobjetivo 21

2.4 A Regra de Decisão 26
2.5 Comentários Finais 28

3 A Curva L e a Curva de Pareto 30
3.1 Problemas Inversos 30
3.1.1 Teoria de Regularização 32
3.1.2 A Curva L 35
3.1.3 A Curva de Validação 38
3.1.4 Simulações Computacionais 41
3.2 Comentários Finais 48

4 Estratégia de Decisão em Classificação 50
4.1 Introdução 50
4.1.1 O Princípio do decisor 51
4.2 O Erro Amostral e o Erro Provável 54
4.3 A Distribuição Binomial 55
4.4 Comentários Finais 56

5 Simulações da Decisão em Classificação 58
5.1 Problema de Duas Distribuições Gaussianas 58
5.1.1 Análise da Probabilidade *a Priori* do Decisor 59
5.1.2 Análise do Desbalanceamento das Classes 71
5.2 Problema de Duas Distribuições em forma de Lua 73
5.2.1 Análise da Sobreposição das Classes 74
5.3 Problema de Diagnóstico do Coração 79
5.3.1 A Base de Dados 80
5.3.2 Experimentos 82
5.3.3 Acurácia dos Classificadores 84
5.4 Comentários Finais 90

6 Estratégia de Decisão em Regressão 91
6.1 A Regra de Decisão por Autocorrelação 91
6.2 Comentários Finais 93

7 Simulações da Decisão em Regressão 95
7.1 Análise de Ruído 95
7.1.1 Problema $f_3(x)$ 96

7.1.2 Análise de Ruído Correlacionado 100
7.1.3 Problema $f_1(x) = sen(x) + \xi$ 104
7.1.4 Problema $f_2(x) = \frac{(x-2)(2x+1)}{(1+x^2)} + \xi$ 108
7.1.5 Problema $f_3(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x}) + \xi$ 112
7.2 Comparação entre decisores 115
7.3 Comentários Finais 117

8 Conclusões e Propostas de Continuidade **118**
Propostas de Continuidade 120

Referências Bibliográficas **121**

Capítulo 1

Introdução

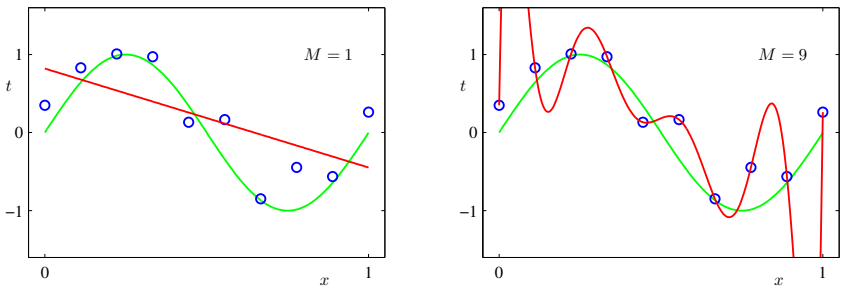
Neste capítulo é apresentada a motivação para a realização deste trabalho de tese, bem como os objetivos principais e específicos. É descrito, sucintamente, o problema para o qual esta tese propõe uma solução estruturada. A organização geral do texto é mostrada em seguida.

1.1 A Teoria do Aprendizado de Máquina

Em problemas de aprendizado de máquina, o objetivo principal é encontrar a função mais adequada para representar uma função geradora desconhecida, dentre um vasto conjunto de possíveis funções, sendo que esta busca é realizada tendo como base um conjunto limitado de dados, ou *exemplos*, que representam a função geradora. Além disso, os exemplos podem estar sujeitos a ruído, o que constitui o caso mais geral (Mitchell, 1997).

O modelo encontrado pode ser uma representação excessivamente simples do conjunto de exemplos, conforme a Figura 1.1a. Por outro lado, o modelo encontrado pode ser uma representação excessivamente complexa do conjunto de exemplos, conforme a Figura 1.1b. O caso mais simples apresenta algum erro em relação aos exemplos disponíveis. Em contrapartida, o modelo mais complexo apresenta erro nulo em relação aos exemplos disponíveis. Mas a questão chave é, qual das alternativas é a solução para o problema em questão?

Em problemas de aprendizagem de máquina, os modelos muito simples inserem uma tendência de polarização nas respostas do modelo. Por outro lado, os modelos muito complexos, por super-estimarem a complexidade do problema, tendem a sofrer de grande variância (ou instabilidade) em suas respostas (Michie et al., 1994).



(a) Underfitting (b) Overfitting

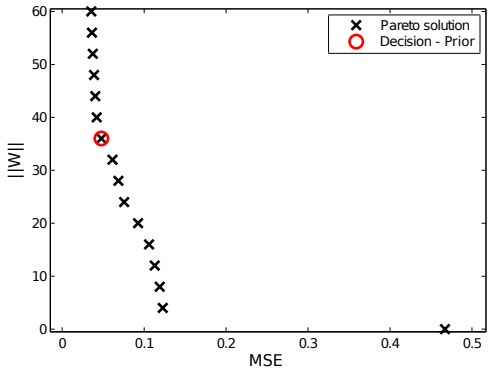
Figura 1.1: Soluções obtidas por modelos excessivamente simples e excessivamente complexos, dado um número limitado de exemplos fornecidos.

Resolver o problema do aprendizado de máquina corresponde ao princípio de se escolher uma função que melhor represente o conjunto de exemplos, de modo que esta função escolhida equilibre os efeitos de polarização e de variância do modelo (Geman et al., 1992) ou minimize o risco estrutural do modelo, como apontado por Vapnik and Chervonenkis (1974). Para resolver este problema existem diversas abordagens que buscam controlar a complexidade da solução.

1.2 O Controle de Complexidade

Geralmente, em problemas de aprendizado supervisionado, o objetivo primal é encontrar um modelo que melhor se aproxime da função geradora (desconhecida) dos exemplos. Como o processo de treinamento do modelo é realizado por meio de um conjunto limitado de exemplos, normalmente corrompido por um tipo de ruído, realiza-se um controle da complexidade do modelo. Este controle é responsável por tornar o modelo de aprendizado suficientemente capaz de aprender a partir dos exemplos e, também limita a capacidade do modelo em aprender o ruído inserido.

Figura 1.2: Exemplo de Conjunto de Soluções Pareto-ótimas e uma escolhida.



Portanto, a resposta da máquina de aprendizagem não deve ter complexidade inferior ao problema, privilegiando uma resposta excessivamente simples (*sub-ajuste*), e nem ser tão complexa a ponto de interpolar todos os exemplos de treinamento (*sobre-ajuste*). Este é o conhecido "dilema entre a polarização e a variância", amplamente discutido no contexto de Redes Neurais Artificiais - RNAs (Geman et al., 1992).

Deste modo, busca-se por uma estrutura de complexidade adequada para um problema, melhorando sua capacidade de generalização à medida que o modelo utilizado é capaz de ser ajustado adequadamente para dados ruidosos.

Neste trabalho, originalmente, o controle de complexidade é realizado por meio da abordagem multiobjetivo do aprendizado supervisionado de RNAs MLP (*Multi-layer Perceptrons*) com estratégias de decisão dependentes de dados de validação (Teixeira et al., 2000). A decisão é tomada por meio de um critério baseado em uma amostra do

conjunto de dados original, que sendo suficiente, tenderá a indicar uma boa estimativa de qual seria o melhor dentre os modelos gerados pelo método multiobjetivo.

1.3 O Problema

Desde que a otimização multiobjetivo vem sendo adotada como uma nova perspectiva de abordagem em aprendizagem de máquina, um novo ponto de vista pôde ser usado sob o controle da complexidade dos modelos. Os métodos baseados em técnicas clássicas de regularização puderam ter uma nova leitura, agora sob a luz a otimização multiobjetivo, permitindo compreender bem seus fundamentos bem consolidados e discutir suas limitações. Neste ponto ao adotar a abordagem multiobjetivo, lidamos com duas condições importantes nesse processo: a representação do conjunto de soluções candidatas e a seleção do modelo. A primeira etapa possui ampla quantidade de estudos de qualidade. No entanto, a seleção de modelos multiobjetivos ainda carecia de estudos mais formais para ampliar o *framework* da aprendizagem de máquina multiobjetivo.

Considerando o cenário com um limitado conjunto de dados, o processo de decisão de referência ainda era o trabalho de Teixeira (2002), capaz de selecionar um modelo que estime o risco estrutural mínimo. Porém há um cenário com uma escassez maior, onde o conjunto de dados é ainda mais reduzido para representar a informação do problema ou até mesmo não se é capaz de afirmar se o conjunto de dados é suficiente. Então, temos uma condição problemática de ser estimar o risco estrutural mínimo de uma máquina de aprendizagem dentro da abordagem multiobjetivo com um conjunto de dados muito reduzido.

Este cenário problemático é o que define o tema central que este trabalho de tese se propôs a estudar e buscar implementar uma estrutura de solução mais geral para o processo de decisão por modelos multiobjetivos em aprendizagem de máquina.

1.4 Objetivo Principal

O método multiobjetivo, desde sua proposta original para redes neurais (denominado MOBJ), baseia-se numa abordagem de obtenção de um conjunto de soluções candidatas, dentre as quais nenhuma é melhor do que a outra segundo o princípio de otimalidade de Pareto (Teixeira et al., 2000). Todavia, o processo de otimização multiobjetivo necessita da indicação de uma solução que represente a solução final que deverá ser implementada. O mesmo acontece trata-se a aprendizagem de máquina sob o ponto de vista multiobjetivo. A escolha da solução final deverá apontar para um modelo que consiga representar um bom equilíbrio entre o ajuste do modelo aos dados e o grau de complexidade desse modelo, equilibrando os efeitos de polarização e variância do aprendizado, discutidos em Geman et al. (1992) e também em Hastie et al. (2001) na especificação do risco estrutural mínimo, de Vapnik and Chervonenkis (1974) .

Dessa forma, o principal objetivo desta tese é formalizar, para o aprendizado multiobjetivo, uma combinação de novas abordagens de tomada de decisão no conjunto

Pareto-ótimo em problemas de aprendizado supervisionado, baseadas em critérios independentes de dados de validação, garantindo ainda uma representação do modelo de risco estrutural mínimo.

1.4.1 Objetivos Específicos

Dentre os resultados esperados deste trabalho, com o intuito de alcançar o objetivo principal, é possível destacar alguns importantes. Um importante objetivo específico deste trabalho é a análise da equivalência do método multiobjetivo com o método de regularização baseado na *Curva L*, descrita em Hansen (2001, 1992); Hansen and O'Leary (1993) na literatura de problemas inversos (Hofmann, 1995) para construção de modelos a partir de dados observados, o que é muito similar ao conceito de aprendizado supervisionado e aproximação de funções (Vito et al., 2005). Este novo olhar permitirá resgatar alguns fundamentos dos métodos de regularização para serem analisados sob uma outra ótica, a da otimização multiobjetivo para aprendizagem supervisionada. Os resultados dos problemas de classificação e regressão deverão fornecer subsídios para uma discussão dessas abordagens.

Além disso, espera-se que a implementação dos métodos de seleção de modelos consiga deixar evidente a vantagem do custo computacional de não se adotar métodos de reamostragem, como o *cross-validation*, no treinamento multiobjetivo. Uma vez que, apesar de serem bons estimadores do erro de generalização e, também independentes da estrutura do modelo, possuem um alto custo de serem implementados.

1.5 Motivação

Em modelos de aprendizado, como as RNAs, Redes de Função de Base Radial (*Radial Basis Function* - RBFs) e Máquinas de Vetores Suporte (*Support Vector Machines* - SVMs), desejamos encontrar uma função adequada para os dados disponíveis durante o treinamento. O problema de aprendizado, seja um problema de regressão ou de classificação, precisa realizar algum controle da complexidade do modelo para que o mesmo responda corretamente para novos exemplos. O método multiobjetivo (MOBJ) surgiu como uma alternativa para a abordagem do dilema entre a polarização e variância em modelos de RNAs MLP (Teixeira et al., 2000). De fato, esta abordagem já obteve resultados que garantem a alta capacidade de generalização dos modelos obtidos através da estratégia de controle de complexidade e que podem ser expandidos para outros modelos de aprendizado, como já foi feito em Carrano et al. (2008); Kokshenev and Braga (2008a) e Suttorp and Igel (2006). Para todos esses modelos estratégias de decisão deveriam ser aplicáveis. Estes foram alguns dos elementos motivadores para dar continuidade ao olhar multiobjetivo da aprendizagem de máquina e suas consequências.

1.6 Justificativa

Dentre os problemas de otimização multiobjetivo, as estratégias de decisão compõem uma parte importante do processo (Parreiras, 2006). Na aprendizagem de máquina multiobjetivo isso não é diferente. No desenvolvimento do método multiobjetivo de treinamento de redes neurais apresentado por Teixeira et al. (2000), a caracterização de uma estratégia de decisão pode variar a cada novo problema, como apresentado nos trabalhos de Teixeira (2002); Teixeira et al. (2007); Medeiros (2007); Torres (2012); Medeiros et al. (2017). Como os trabalhos desenvolvidos nessa linha apresentaram predominantemente a ênfase na etapa de representação do conjunto de soluções candidatas, a decisão era implementada por um método simples baseado no erro de validação, com algumas limitações. Dessa forma, havia um campo a ser explorado com mais atenção. Por esse motivo, decidiu-se explorar de maneira mais aprofundada a etapa de seleção, gerando novas abordagens que permitissem compreender melhor cada problema de aprendizagem.

Além do acima citado, em um cenário geral todas estratégias deveriam convergir para a indicação de uma solução que minimize o risco estrutural (erro de generalização) a partir dessa abordagem multiobjetivo. De acordo com essa abordagem, a seleção de modelos deveria ser capaz de indicar um modelo candidato capaz de apresentar um desempenho satisfatório segundo os critérios definidos para cada problema de aprendizagem. Ao considerarmos um cenário ótimo com um vasto conjunto de dados, o problema de seleção de seleção estaria resolvido. No entanto, o cenário real é frequentemente apresentado na forma de um conjunto limitado de dados (Andonie, 2010; Shaikhina and Khovanova, 2017). Portanto, o método de seleção baseado na separação dos dados em conjunto de treinamento, validação e de teste não garante a qualidade dos conjuntos na representação da informação a ser aprendida. Dessa forma, os métodos de seleção de modelos por reamostragem poderão gerar duas condições indesejadas durante a abordagem multiobjetivo: a amostra de validação pode não ser suficiente ou a técnica de reamostragem pode ser excessivamente custosa quando associada à abordagem multiobjetivo, que também já é custosa.

Após os pontos mencionados, a função deste trabalho de tese é propor e implementar métodos para seleção de modelos no conjunto de soluções Pareto-ótimas obtido segundo os critérios estabelecidos para a natureza do problema de aprendizagem em particular. Em todos os casos, o modelo escolhido deverá comportar-se de maneira adequada à natureza do problema estudado e garantindo que o modelo escolhido seja uma representação válida do risco estrutural mínimo. Além disso, este trabalho apresentará formas de se implementar a seleção de modelos usando informações extraídas do treinamento e usando informações conhecidas sobre os dados do problema para permitir que a seleção adote um critério bem definido de seleção.

1.7 Contribuições

A principal contribuição desta tese está na formalização de uma estrutura de decisão multiobjetivo baseada em informações complementares sobre o problema de aprendiza-

gem, possibilitando a dispensa do uso de técnicas de reamostragem dos dados disponíveis. Esta abordagem garante um mecanismo de seleção baseado em critérios bem definidos e mais consistente do que os métodos de seleção por erro de validação, adotados amplamente no treinamento multiobjetivo de redes neurais.

A abordagem de decisão em problemas de regressão busca pela solução que apresente o resíduo de mínima auto-correlação, conforme o trabalho desenvolvido em Medeiros (2007). Já a abordagem de decisão em problemas de classificação busca pela solução usando um método de seleção baseado num critério de teste de hipótese, desenvolvido em Medeiros et al. (2017). Essas duas abordagens permitiram um olhar mais amplo sobre o problema da seleção de modelos mutiobjetivos, visando a independência do conjunto de dados de validação. Além disso, os resultados dos trabalhos citados garantiram a boa estimativa do erro de generalização dos modelos escolhidos nos casos estudados.

Uma outra contribuição desse trabalho foi abordar a representação das soluções Pareto-ótimas sob um olhar particular bem conhecido nos métodos de regularização. O método da *Curva L*, proposto por Hansen (1990) em métodos de regularização para problemas inversos representa soluções com complexidades controladas e uma região indicativa da solução que equilibra os efeitos de *overfitting* e *underfitting*. Neste ponto a visão multiobjetivo da aprendizagem permitiu uma nova análise dos aspectos particulares da abordagem da *Curva L* no controle da complexidade em problemas de aprendizagem.

1.8 Organização do Texto

Esta tese é organizada da seguinte maneira:

- O Capítulo 2 apresenta os resultados do método multiobjetivo de aprendizado aplicado ao treinamento de redes neurais artificiais em problemas de aprendizado. Os principais conceitos relacionados ao assunto são descritos, bem como os resultados de experimentos computacionais que destacam a aplicação do método de treinamento com as novas estratégias de tomada de decisão já propostas e implementadas até o momento para o método MOBJ.

- O Capítulo 3 apresenta o método de regularização apoiado pelo critério da *Curva L* e uma análise comparativa deste critério com o método multi-objetivo de aprendizado. Serão descritos os conceitos relacionados ao assunto e experimentos computacionais que irão ressaltar as semelhanças e diferenças entre os métodos. As semelhanças, diferenças e limitações entre as abordagens serão os principais pontos deste capítulo. A abordagem multiobjetivo mostrará ter um espectro de aplicação maior que a *Curva L* e garantir que melhores resultados podem ser atingidos.

- O Capítulo 4 apresenta a estratégia de decisão em problemas de Classificação de Padrões, incorporando no decisor um teste estatístico de hipóteses.

- O Capítulo 5 apresenta os resultados dos experimentos computacionais das diferentes estratégias de decisão adotadas no método MOBJ em classificação de padrões.

Os gráficos desse capítulo apresentam a solução ótima de acordo com cada estratégia e sua sensibilidade a variações nas condições de ajuste do modelo, como por exemplo, o nível de incerteza inserido no conjunto de pontos amostrados.

- O Capítulo 6 apresenta a estratégia de decisão em problemas de regressão, baseado na autocorrelação do resíduo do modelo.
- O Capítulo 7 apresenta os resultados dos experimentos computacionais das diferentes estratégias de decisão adotadas no método MOBJ em problemas de regressão. Os resultados desse capítulo apresentam a solução ótima de acordo com cada estratégia e sua sensibilidade a variações nas condições de ajuste do modelo, como por exemplo, o nível de ruído inserido no conjunto de pontos amostrados.
- No Capítulo 8 são apresentadas as Conclusões e Propostas de Continuidade deste trabalho.

Capítulo 2

Decisão Multiobjetivo

2.1 Aprendizado de Máquina e a Generalização

2.1.1 Princípios do Aprendizado de Máquina

A aprendizagem de máquina é um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa por meio de sua interação com o meio externo (Mitchell, 1997). Dentre as formas de aprendizagem conhecidas, este trabalho está voltado para o problema do aprendizado supervisionado. O aprendizado supervisionado é caracterizado pela presença de um supervisor. O supervisor detém o conhecimento do problema e o conjunto de dados é a representação deste conhecimento.

No aprendizado supervisionado busca-se modelar a relação entre os dados de entrada e a respectiva saída do conjunto de exemplos, descartando o possível ruído presente nos dados. Existem infinitas relações que podem ser obtidas por meio do conjunto de dados fornecido. A questão a ser respondida é: *"Qual a relação mais apropriada?"*, ou seja, *"Qual a máquina de aprendizagem que melhor descreve (aproxima) a função geradora dos dados?"*. A máquina de aprendizagem adequada apresentará alta precisão para novos exemplos.

Assim, denomina-se capacidade de generalização a habilidade de uma máquina em dar respostas coerentes para dados novos (Braga et al., 2000). Entretanto, a minimização do erro da generalização não é uma característica inerente ao processo de treinamento - é preciso inserir uma forma de controle da complexidade da máquina de aprendizagem para que seja modelada apenas a informação útil. A Figura 2.1 mostra o esquema de um sistema de aprendizado.

A Figura 2.1 apresenta os três componentes básicos de um sistema de aprendizado de máquina supervisionado. São eles:

- Um gerador de exemplos x, amostrados independentemente a partir de uma distribuição de probabilidade fixa, porém desconhecida.

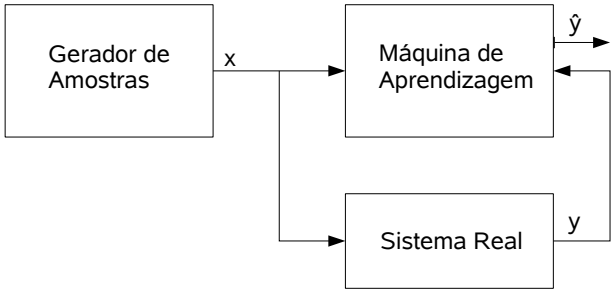


Figura 2.1: O Aprendizado de Máquina Supervisionado.

- Um sistema real (supervisor) que retorna uma saída y para uma dada entrada x, de acordo com uma função de distribuição condicional $P(y|x)$, também fixa mas desconhecida.
- Uma máquina de aprendizagem capaz de realizar um conjunto de funções $\hat{y} = f(x; w), w \in \mathbf{W}$ no espaço de hipóteses, composto por funções ou pelos parâmetros delas (w).

Portanto, pode-se definir o problema de aprendizado como o problema de escolher, em um dado conjunto de funções $(f(x; w), w \in \mathbf{W})$, aquela que melhor se aproximar da resposta do sistema real. A medida da discrepância (ou perda) entre a resposta do sistema real e a resposta gerada pela máquina de aprendizagem pode ser dada pelo funcional de risco (ou erro) (Vapnik, 1998):

$$R(w) = \int L(y, f(x; w))dF(x, y), \tag{2.1}$$

onde $L(y, f(x; w))$ é um funcional de perda qualquer e $R(w)$ é o valor esperado do erro de teste (ou generalização).

No entanto, para encontrar a função que minimiza o funcional de risco, a única informação disponível sobre o problema são as amostras do conjunto de dados. Dessa forma, os problemas de aprendizado, sejam eles de regressão ou classificação, são dependentes das amostras disponíveis. Os problemas de classificação e de regressão são, então, particularidades do caso geral do funcional de risco, dado na Equação 2.1.

Problema de Regressão: *"Um problema de regressão pode ser definido da seguinte forma: "Minimizar o funcional de risco da Equação 2.1 segundo a função de perda:*

$$L(y, f(x; w)) = (y - f(x; w))^2 \tag{2.2}$$

onde a resposta do sistema real (y) é dada por um valor real."

Problema de Classificação: *"Um problema de classificação pode ser definido da seguinte forma: "Minimizar o funcional de risco da Equação 2.1 segundo a função de perda:*

$$L(y, f(x; w)) = \begin{cases} 0 \ se \ y = f(x; w) \\ 1 \ se \ y \neq f(x; w) \end{cases} \tag{2.3}$$

onde a resposta do sistema real pode ter somente dois valores $y \in \{0, 1\}$ e $f(x; w)$ é uma função indicadora, pois só pode representar valores zero ou um. Para essa função de perda, o funcional de risco indica a quantidade de classificações incorretas."

O princípio usado na regressão, como também na classificação, é chamado de princípio de minimização do risco empírico (Vapnik, 1998). No entanto, somente a minimização desse funcional de risco sobre o conjunto de amostras disponíveis nem sempre é suficiente para a obtenção de respostas precisas em novos exemplos, pois ela não leva em consideração a complexidade do sistema real. Para garantir que a máquina de aprendizagem fornecerá uma aproximação razoável da função geradora dos exemplos, o algoritmo de aprendizado precisa implementar algum controle da complexidade das funções $(f(x; w), w \in \mathbf{W})$.

2.1.2 Capacidade de Generalização

A capacidade de generalização de uma máquina de aprendizagem está sujeita à qualidade da informação disponível sobre o problema (o conjunto de dados), e além disso ao método de controle da complexidade da função aproximada que tiver sido empregado (Hastie et al., 2001).

O conjunto de exemplos usado é fundamental para definir a qualidade da solução obtida. Entretanto, um dos fatores complicadores do aprendizado é, exatamente, a disponibilidade de exemplos suficientes para representar toda informação do problema em particular. Esse problema torna-se ainda mais crítico quando o número de exemplos necessário para um bom desempenho de uma máquina de aprendizagem cresce exponencialmente com o número de dimensões. Esse problema é denominado, na literatura, como *Maldição da Dimensionalidade* e descrito em Bellman (1961).

Para se obter soluções com alta precisão para um conjunto de dados de teste (alta capacidade de generalização) é sempre bem vindo o uso de alguma informação prévia. Essa informação, quando disponível, carrega consigo um conhecimento sobre a natureza do problema e pode ajudar na obtenção de uma solução melhor. Entretanto, nem sempre a informação prévia está disponível, nem sempre é suficiente ou até mesmo quando há, é possível que não se saiba como usá-la, tornando o aprendizado uma tarefa mais complexa. Em princípio, a própria teoria da regularização funciona como uma abordagem de inserção de conhecimento prévio para realizar o controle da complexidade da solução obtida pelo aprendizado (Tikhonov and Arsenin, 1977).

Na teoria do aprendizado estatístico apresentada em Vapnik (1998), utiliza-se o princípio de minimização do risco estrutural (SRM - *Structural Risk Minimization*), definido

em Vapnik and Chervonenkis (1974) para selecionar uma função desejada no espaço de hipóteses. Este é um princípio indutivo para seleção de modelos a partir de um conjunto dados finito. Ele descreve um modelo geral de controle da capacidade e fornece um *trade-off* entre a complexidade do espaço de hipóteses (a dimensão VC das hipóteses) e qualidade do ajuste nos dados de treinamento (erro empírico). O procedimento é destacado abaixo:

1. Usando conhecimento a priori do domínio, escolha uma classe de funções. Por exemplo, polinômios de grau p, redes neurais com n neurônios de camada oculta.

2. Divida a classe de funções em uma hierarquia de subconjuntos aninhados em ordem crescente de complexidade. Por exemplo, polinômios de grau crescente.

3. Execute a minimização do risco empírico em cada subconjunto (essa é essencialmente a seleção de parâmetros).

4. Selecione o modelo cuja soma do risco empírico e dimensão VC (complexidade) é mínima.

Diferentes abordagens como o early stopping (Weigend et al., 1990), a maximização da margem de separação (Cortes and Vapnik, 1995; Smola et al., 1999), os métodos de regularização (Tikhonov and Arsenin, 1977; Weigend et al., 1990; Girosi et al., 1995; Burden and Winkler, 2009), visam minimizar direta ou indiretamente, o risco estrutural em máquinas de aprendizagem como MLPs, RBFs e SVMs. A abordagem multiobjetivo do problema de aprendizagem de máquinas para implementar o princípio SRM define uma nova ótica na natureza do problema (Braga et al., 2006) .

2.1.3 Seleção de Modelos

Como parte do processo de obtenção do modelo de melhor ajuste aos dados, é necessário uma abordagem que determine sistematicamente o modelo adequado dentre diversos modelos propostos. A seleção de modelos consiste neste processo. Há alguns anos diversos trabalhos tem sido desenvolvidos no intuito de encontrar abordagens para indicação de um modelo dentre um conjunto de modelos candidatos, como é destacado no trabalho de revisão de Oneto (2018) que aborda uma visão geral dos problemas de seleção de modelos com base na teoria do aprendizado estatístico de modo a torná-la mais acessível e utilizável na prática. No trabalho de revisão de Raschka (2018) foram destacados os métodos de reamostragem para seleção de modelos e suas limitações para amostras de tamanho reduzido, sugerindo métodos alternativos para tratar o problema da seleção em amostras reduzidas.

A seleção de modelos é o processo de escolha de um modelo final de aprendizado de máquina dentre um conjunto de modelos candidatos a partir de um conjunto de dados de treinamento. Este processo pode ser aplicado à diferentes tipos de modelos como, por exemplo, regressão logística, RNAs, RBFs, SVMs, KNN, etc (Ding et al., 2018). Também pode ser aplicado em modelos do mesmo tipo mas com configurações diferentes de seus

hiper-parâmetros como, por exemplo, o tipo de função de kernel de uma SVM (Gold and Sollich, 2003).

Em situações práticas há, rotineiramente, o desafio de se ajustar um modelo preditivo de regressão ou de classificação a partir de um conjunto de dados. Porém, de antemão, é desconhecido qual é o modelo que melhor se ajusta aos dados. Nesse caso, estabelece-se um conjunto de diferentes modelos candidatos para que seja feito o ajuste e a avaliação do desempenho desses candidatos com o objetivo de identificar um modelo final.

Para compreender melhor o desafio do processo de seleção de modelos, é preciso compreender a ideia do que seja o "melhor modelo". Todos os modelos apresentam algum erro preditivo, dado pelo ruído nos dados, pela incompletude da amostra de dados e pelas limitações de cada tipo diferente de modelo adotado. Dessa forma, a noção do melhor modelo ou perfeito, não é muito útil. Ao invés disso, a busca deve-se concentrar na obtenção de um modelo "bom o suficiente".

A identificação do modelo que atenda aos interesses do projeto precisa conhecer os requisitos requeridos pelos projetistas para ajudar no direcionamento do processo de seleção. Em alguns projetos, o limite da complexidade do modelo pode ser um requisito primordial e neste caso é preferido um modelo de menor habilidade de ajuste aos dados, mas mais simples e fácil de compreender. Em contrapartida, em outros projetos, a habilidade do modelo se ajustar aos dados pode ser preferida independentemente da complexidade. Portanto, o "bom o suficiente"pode ser diferente de projeto para projeto, e para isso deverá ser compreendido o contexto como um todo.

Técnicas para Seleção de Modelos

A melhor abordagem para a seleção de modelos requer dados suficientes, podendo ser praticamente infinitos, dependendo da complexidade do problema.

Nessa situação ideal, dividem-se os dados em conjuntos de treinamento, validação e teste, em seguida, ajustam-se os modelos candidatos no conjunto de treinamento, avaliam-se e selecionam-se os modelos no conjunto de validação, sendo reportado o desempenho do modelo final no conjunto de teste. Conforme Hastie et al. (2001), em um cenário rico de dados, a melhor abordagem é dividir aleatoriamente o conjunto de dados em três partes: um conjunto de treinamento, um conjunto de validação e um conjunto de testes. O conjunto de treinamento é usado para ajustar os modelos; o conjunto de validação é usado para estimar o erro de previsão para a seleção do modelo; o conjunto de testes é usado para avaliar o erro de generalização do modelo final escolhido.

No entanto é impraticável na maioria dos problemas de modelagem preditiva, uma vez que raramente há dados suficientes ou não há capacidade de julgar o que seria suficiente. Em muitas aplicações, no entanto, o fornecimento de dados é limitado e, para obter bons modelos, é desejável usar o máximo de dados disponíveis para o treinamento. No entanto, se o conjunto de validação for pequeno, ele fornecerá uma estimativa relativamente ruidosa do desempenho preditivo.

Para abordar esse problema, duas classes de métodos podem ser estabelecidas para a seleção de modelos. São elas:

- Métodos Analíticos
- Métdos de Reamostragem

Os **métodos analíticos** envolvem uma avaliação quantitativa dos modelos candidatos usando o desempenho no conjunto de treinamento e a complexidade do modelo. Sabe-se que o erro de treinamento é uma estimativa otimista e, portanto, não é uma boa base para a escolha de um modelo. O desempenho esperado do modelo pode ser penalizado com base no quão otimista se acredita que o erro de treinamento seja. Isso geralmente é alcançado usando métodos específicos, geralmente lineares, que penalizam modelos complexos. Um modelo com poucos parâmetros é menos complexo, e por causa disso, é preferido porque na média ele generaliza melhor. Na literatura pode-se encontrar diversas métricas analíticas de seleção de modelos. É possível destacar três delas:

Akaike Information Criteria (AIC): A métrica AIC está fundamentada na teoria da informação (Akaike, 1974) . Quando um modelo estatístico é usado para representar o processo que gerou os dados, a representação quase nunca será exata. Algumas informações serão perdidas usando o modelo e a AIC estima a quantidade relativa de informações perdidas por um determinado modelo: quanto menos informações um modelo perde, maior a qualidade desse modelo. A métrica AIC dá maior importância ao desempenho do modelo no conjunto de treinamento do que na complexidade do modelo. Logo, a seleção baseada na AIC apresenta a tendência de privilegiar modelos mais complexos (Murphy, 2013).

Bayesian Information Criteria (BIC): A métrica BIC, proposta por Schwarz (1978), foi definida em termos da probabilidade *a posteriori*. Tal como a AIC é apropriada para modelos que se enquadram no framework de estimativa da máxima verossimilhança. A métrica entrega valores diferentes da AIC, embora proporcionais. A métrica BIC, comparada com a AIC, penaliza mais rigorosamente os modelos complexos, que significa que esses terão menor probabilidade de serem selecionados (Weakliem, 1999).

Minimum Information Length (MDL): A métrica MDL surgiu no campo de estudo da teoria da informação no trabalho de Rissanen (1978). A teoria da informação está preocupada com a representação e transmissão de informações em um canal ruidoso e, como tal, mede quantidades como entropia, que é o número médio de bits necessários para representar um evento a partir de uma variável aleatória ou distribuição de probabilidade (MacKay, 2003; Witten et al., 2011). A métrica representa o número de bits necessários para descrever os dados e o modelo, sendo que a seleção do modelo é feita pelo modelo que minimiza a soma dessas duas descrições.

Essas métricas são apropriadas quando usa-se modelos lineares simples como regressão linear, regressão logística, onde o cálculo da penalidade da complexidade do modelo é conhecido e tratável (Neath and Cavanaugh, 2012; Piironen and Vehtari, 2017). No

entanto, alguns trabalhos recentes apresentaram resultados novos que permitem uso das métricas baseadas em critério de informação para seleção de modelos não-lineares, como foi apresentado em Gu et al. (2018) com o APRESS (*Adjustable Prediction Error Sum of Squares*). Há ainda diversas outras métricas como mostra Ding et al. (2018) mas fogem do escopo desse trabalho.

Os **métodos de reamostragem** envolvem a estimação do desempenho do modelo em dados fora do conjunto de treinamento. Isto é alcançado através da divisão do conjunto de treinamento em subconjuntos de treinamento e testes, ajustando os modelos no subconjunto de treinamento e avaliando no subconjunto de teste. Esse processo pode se realizar múltiplas vezes e o desempenho médio alcançado em cada execução é reportado. Estes métodos são bem gerais pois não requerem nenhuma suposição paramétrica dos modelos candidatos. O trabalho de Xu and Goodacre (2018) avaliou e discutiu os resultados de alguns métodos em um estudo comparativo, dos quais nos concentramos nos mais comuns.

Dos métodos mais conhecidos de seleção de modelos por reamostragem temos:

Método *Holdout* (ou Validação): Este método particiona os dados em dois subconjuntos mutuamente exclusivos: um conjunto de treinamento e um conjunto de teste. A proporção frequentemente adotada é a de $2/3$ do total para o treinamento e o restante para teste. O conjunto de treinamento é fornecido ao algoritmo de aprendizagem e o classificador induzido é avaliado com o conjunto de teste e o erro de predição calculado. Assumindo que a acurácia do classificador aumenta à medida que mais dados são vistos, o método *holdout* é um estimador pessimista porque somente uma porção dos dados é fornecida ao modelo para treinamento. Quanto mais dados forem deixados para o conjunto de teste, maior o bias da estimativa do modelo; no entanto menos dados para o conjunto de teste significa que o intervalo de confiança para a acurácia será maior (Raschka, 2018).

Cada dado de teste pode ser visto como uma tentativa de Bernoulli: uma predição é correta ou incorreta. Considerando S o número de classificações corretas no conjunto de teste, então S apresenta uma distribuição binomial (soma das tentativas de Bernoulli). Para conjuntos de teste razoavelmente grandes, a distribuição S/N_t (N_t é o número de dados de teste) é aproximadamente normal com média acc e variância $acc \cdot (1 - acc)/N_t$, segundo o trabalho de Raschka (2018), onde acc é a referência da medida de acurácia do modelo.

A estimativa do *holdout* é um número aleatório que depende da divisão em um conjunto de treinamento e um conjunto de testes. Em uma sub-amostragem aleatória, o método *holdout* é repetido K vezes e a acurácia estimada é derivada da média das execuções. O desvio padrão pode ser estimado como o desvio padrão das estimativas de acurácia de cada execução do *holdout*.

A principal suposição que é violada na sub-amostragem aleatória é a independência dos exemplos no conjunto de testes daqueles no conjunto de treinamento. Se o conjunto de treinamento e teste for formado por uma divisão a partir de um conjunto de dados original, uma classe sobre-representada em um subconjunto será

sub-representada no outro. Em linhas gerais o método *holdout* faz uso ineficiente dos dados disponíveis.

***Cross-Validation* :** No *k-fold cross-validation* o conjunto de dados original é aleatoriamente particionado em k subconjuntos de tamanhos iguais (ou quase iguais em alguns casos). Desses k subconjuntos, um é retido para a validação e os $k-1$ restantes são usados para o treinamento. O processo de *cross-validation* é então repetido k vezes usando um subconjunto uma única vez como o conjunto de validação. A estimativa de acurácia geral é dada pelo número total de classificações corretas dividido pelo número de exemplos da conjunto de dados completo. A vantagem deste método em relação ao *holdout* é que todos exemplos são usados para treinamento e validação, e cada exemplo é usado para validação exatamente uma vez. O 10-fold *cross validation* é frequentemente usado, mas em geral o k é um parâmetro a ser escolhido.

No *stratified cross-validation* as partições são selecionadas para que elas contenham aproximadamente a mesma proporção dos rótulos do conjunto de dados original. No caso de classificação binária, isso significa que cada partição contém aproximadamente as mesmas proporções dos dois tipos de rótulos de classe.

No *leave-one-out cross-validation*, um caso particular onde $k = n$, em que n é o número de exemplos do conjunto de dados original, cada partição de teste possui apenas um exemplo.

Esses e outros métodos de *cross-validation* caracterizam-se por exigir um número maior execuções dos algoritmos de aprendizagem para obter a estimativa de erro de predição para a seleção do modelo. Isso pode não ser desejado em alguns casos, mas funcionam muito bem para conjunto de dados reduzidos. O método *cross-validation* visa fornecer estabilidade às estimativas dos modelos, por meio do uso eficiente do conjunto de dados.

Bootstrap: A família Bootstrap foi introduzida por Efron e é totalmente descrita em Efron and Tibshirani (1993). Dado um conjunto de dados de tamanho n, uma amostra bootstrap é criada pela amostragem de n instâncias uniformemente a partir dos dados (com substituição). Como o conjunto de dados é amostrado com substituição, a probabilidade de qualquer instância não ser escolhida após n amostras é $(1-1/n)^n \approx \exp^{-1} \approx 0.368$; o número esperado de instâncias distintas do conjunto de dados original que aparece no conjunto de testes é $0.632n$. A estimativa de acurácia $e0$ é obtida usando a amostra bootstrap para treinamento e o restante das instâncias para teste. Dado um número b, o número de amostras bootstrap, seja $e0_i$ a estimativa de precisão para a amostra bootstrap i. A estimativa bootstrap 0.632 é definida como:

$$acc_s = \frac{1}{b}\sum_{i=1}^{b}(0.632 \cdot e0_i + 0.368 \cdot acc_s)$$

onde acc_s é a estimativa de acurácia com resubstituição sobre o conjunto de dados completo. A suposição feita pelo bootstrap é basicamente a mesma da *cross-validation*, ou seja, a estabilidade do algoritmo sobre o conjunto de dados.

Há diversos cenários que os métodos analíticos de seleção de modelos não estão disponíveis para cálculo, em alguns o desafio é o tipo de problema: classificação ou regressão; em outros o desafio é o tipo de modelo adotado. Por isso os métodos de reamostragem são frequentemente mais usados para estimação do erro de predição, por serem mais generalistas. E dentre os métodos de reamostragem, de longe, o mais popular são as variação da *cross-validation*. Em contra-partida, a família *cross-validation* carrega consigo o peso de execuções repetidas do algoritmo de treinamento. No contexto deste trabalho de pesquisa a escolha do modelo adequado seguia inicialmente o princípio do método *holdout* (Teixeira, 2002). A abordagem baseada em critério de informação AIC e BIC no aprendizado multiobjetivo foi proposta e desenvolvida no trabalho de Kokshenev and Braga (2010), porém em modelos lineares, neste caso as Redes Neurais de Função de Base Radial. Então esse é o contexto no qual está inserido o objetivo da proposta e desenvolvimento de abordagens alternativas de seleção de modelos na perspectiva do treinamento multiobjetivo de redes neurais do tipo MLP.

2.1.4 Redes Neurais MLP

Neste trabalho, adotou-se o uso das redes neurais *multi-layer* perceptrons (MLPs) para avaliar os métodos propostos. As redes MLPs representam um modelo de máquina de aprendizagem amplamente utilizado na literatura e, também, um dos modelos usados para implementação de algoritmos de treinamento multiobjetivo.

O passo fundamental para o surgimento das RNAs foi o modelo matemático de um neurônio proposto em McCulloch and Pitts (1943). O modelo denominado MCP (McCulloch e Pitts) apresenta n entradas em que cada entrada é multiplicada por um determinado peso w e, em seguida, os resultados são somados e comparados a um limiar u, conforme descrito na Equação 2.4. A função $\phi(\cdot)$ é uma função de ativação do tipo limiar, com resposta $+1$ quando $\sum_{j=1}^{n} w_j x_j - u > 0$ e com resposta em 0, caso contrário. Esse modelo é amplamente citado na literatura de RNAs (Braga et al., 2000).

$$\hat{f} = \phi(\sum_{j=1}^{n} w_j x_j - u) \tag{2.4}$$

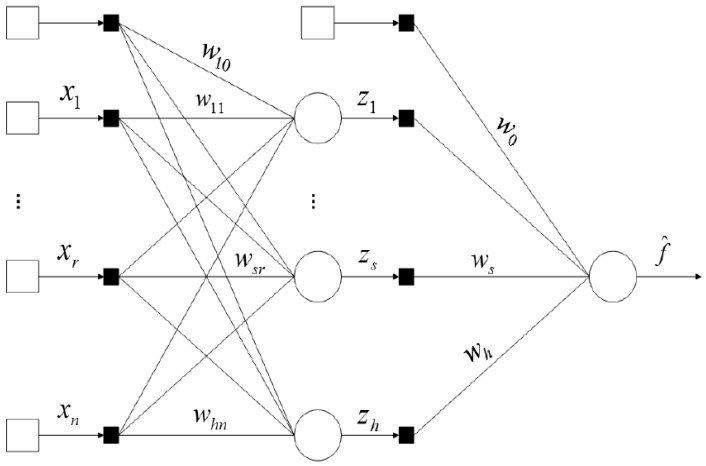
Os pesos positivos estão associados a sinapses *excitatórias* enquanto os pesos negativos modelam as *inibidoras*. O modelo de McCulloch-Pitts foi generalizado de diversas maneiras, uma delas foi usar diversas funções de ativação diferentes da função de limiar. A função de ativação sigmoide é uma das mais usadas. Ela é um função estritamente crescente exibindo suavidade e propriedades assintóticas. A função logística é uma função sigmoide padrão, definida na Equação 2.5.

$$g(x) = \frac{1}{1 + \exp^{(-\beta x)}} \tag{2.5}$$

onde β é um parâmetro de inclinação.

Conhecida essa estrutura, a rede MLP pode ser modelada com um conjunto de n entradas, uma camada escondida com h neurônios e uma camada de saída contendo uma única unidade, conforme ilustrada na Figura 2.2.

Figura 2.2: Topologia de uma Rede Multi-Layer Perceptron de uma camada oculta.



O valor de saída obtido na unidade escondida s da rede, como consequência da apresentação de um vetor de entrada $x = x_1, x_2, \cdots, x_n$ é dado pela seguinte expressão,

$$z_s = \phi(u_s) = \phi(\sum_{r=0}^{n} w_{sr} x_r) \tag{2.6}$$

onde w_{sr} representa um peso entre o neurônio oculto s e a entrada s; $\phi(\cdot)$ é uma função de ativação. De modo equivalente, o valor obtido no neurônio de saída da rede é calculado com base nas saídas emitidas pelos neurônios escondidos,

$$\hat{f} = \phi(v) = \phi(\sum_{s=0}^{h} w_s z_s) \tag{2.7}$$

na qual w_s representa um peso entre o neurônio de saída e o neurônio oculto s. O termo de *bias* foi considerado como um neurônio extra com valor igual a 1. A resposta final da rede MLP para o exemplo x é a saída $\hat{f}$.

A seguir será descrita a formalização de um problema de otimização multiobjetivo e, posteriormente, da aprendizagem de máquina multiobjetivo.

2.2 A Otimização Multiobjetivo

Um problema de otimização multiobjetivo parte da constatação de que, na existência de múltiplos objetivos, existirão soluções que admitirão com que todos os objetivos

melhorem simultaneamente, implicando na existência de soluções melhores. Ainda assim existirão outras soluções para as quais não será possível a melhoria simultânea de todos objetivos, sendo que a melhora em um objetivo implicará na degradação de outro objetivo. Definir a região onde está presente este último conjunto de soluções, denominadas soluções Pareto-ótimas (PO) ou não-dominadas (Pareto, 1896), é um principais problemas da otimização multiobjetivo (Sawaragi et al., 1985; Ferreira, 1999). A seguir, na Equação 2.8 é representada a formulação matemática básica de um problema de otimização multiobjetivo:

$$\begin{aligned} &\text{Max/Min } f_m(x), \quad m = 1, 2, \cdots, M \\ &\text{sujeito a: } \begin{cases} g_j(x) \geq 0, & j = 1, 2, \cdots, J \\ h_k(x) = 0, & k = 1, 2, \cdots, K \\ x_i^{(L)} \leq x_i \leq x_i^{(U)}, & i = 1, 2, \cdots, n \end{cases} \end{aligned} \tag{2.8}$$

onde $f_m(x)$ correspondem às M funções objetivo que deseja-se otimizar, seja na forma de maximização ou de minimização, x é o vetor de n variáveis de decisão $x = (x_1, x_2, \cdots, x_n)^T$, J e K indicam o número de restrições de desigualdade e igualdade, respectivamente. Os valores $x_i^{(L)}$ e $x_i^{(U)}$, representam os limites mínimo ((L)) e máximo ((U)) para a variável x_i.

2.2.1 O Conceito Pareto - ótimo

A falta de uma solução ótima global que atenda todos os objetivos simultaneamente faz com que os problemas de otimização multiobjetivo sejam caracterizados por um conjunto de soluções. Este conjunto pode ser definido usando-se as relações de dominância para caracterizar as soluções não-dominadas, que representam o *trade-off* entre os objetivos do problema. A seguir, usando como exemplo um problema de minimização, serão destacados os conceitos de dominância e de otimalidade de Pareto.

A Relação de Dominância

Em um problema de minimização, dados dois vetores de soluções x_A e $x_B \in \Re^n$ pode-se afirmar que x_A domina x_B se em pelo menos uma dimensão i de um total de m dimensões, $f_i(x_A)$ for estritamente menor que $f_i(x_B)$ e nas demais dimensões $j \neq i$, x_A for menor ou igual a x_B. Assim pode-se afirmar, nesse caso, que o vetor x_A é dominante ou superior enquanto o vetor x_B é considerado dominado ou inferior. Com base nessa definição tem-se a descrição formal dessa relação:

$$\begin{aligned} x_A \prec x_B \leftrightarrow \forall j \in \{1, 2, \ldots, m\},\ f_j(x_A) \leq f_j(x_B) \wedge \\ \exists i \in \{1, 2, \ldots, m\},\ f_i(x_A) < f_i(x_B). \end{aligned} \tag{2.9}$$

Caso não seja possível estabelecer uma definição sobre a relação de dominância entre duas soluções x_A e x_B, pode-se afirmar que x_A é indiferente a x_B.

A Solução Pareto-ótima

Conhecido o conceito da relação de dominância é possível introduzir a condição de otimalidade para problemas de otimização multiobjetivo. Definindo-se a região de soluções viáveis Ω para o problema multiobjetivo em questão, cada possível solução $x_A \in \Omega$ é considerada uma solução Pareto-ótima se não existe nenhuma outra $x_B \in \Omega$ tal que $x_B \prec x_A$. As soluções Pareto-ótimas são também conhecidas como soluções não-inferiores ou não-dominadas. Desse modo, pode-se estender a definição da solução Pareto-ótima para uma região.

Fronteira Pareto-ótima

O critério de otimalidade para problemas de otimização multiobjetivo define então um conjunto de soluções Pareto-ótimas, descrito como:

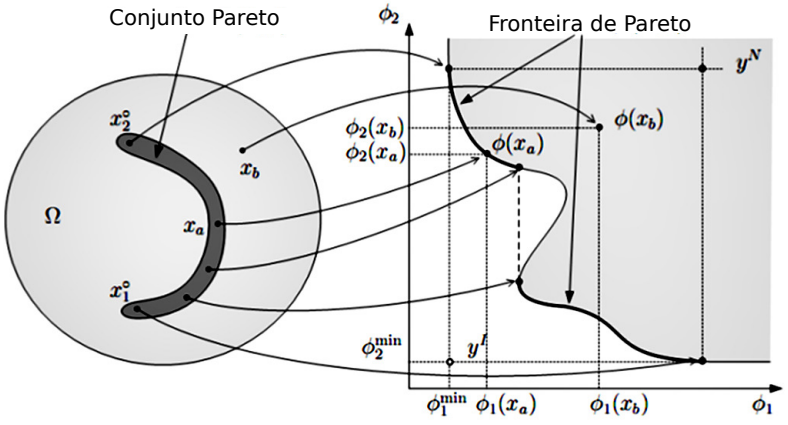
$$P^* := \{x_A \in \Omega | \neg \exists x_B \in \Omega : x_B \prec x_A\}. \tag{2.10}$$

onde $F(x) = [f_1(x), f_2(x), \ldots, f_m(x)]$.

A este conjunto P^* corresponde um conjunto de vetores de avaliações das funções objetivo, constituindo no espaço dos objetivos uma fronteira Pareto-ótima ou, simplesmente, Fronteira de Pareto.

$$FP^* := \{F(x) | x \in P^*\} \tag{2.11}$$

Figura 2.3: Um Exemplo particular de um mapeamento das soluções no plano dos objetivos, onde a região de soluções não-dominadas está destacada em negrito.



A Figura 2.3 ilustra o mapeamento de soluções $x_I \in \Omega$ em soluções $\phi_j(x_i)$ agora representadas no plano dos objetivos $\phi_j(\cdot)$, onde j é o índice de referência do j-ésimo objetivo. A fronteira de Pareto está destacada para mostrar a região correspondente à localização das soluções não-dominadas.

O Processo de Decisão no Conjunto Pareto

O problema de otimização multiobjetivo não encontra-se resolvido com a constituição do conjunto de soluções Pareto-ótimas. Cabe a um decisor a tarefa de escolher a solução final que será implementada. Essa tarefa é o processo de decisão no conjunto Pareto, que pode ser estudada com maiores detalhes em Triantaphyllou (2000) e Parreiras (2006).

A tarefa de decisão usa um critério de avaliação das soluções não mensurável durante o processo de otimização para comparar as soluções entre si e escolher a melhor segundo esse critério. O decisor pode ainda, se necessário, usar informações extras para ajudar na tarefa de decisão.

2.3 O Aprendizado de Máquina Multiobjetivo

A Otimização multiobjetivo tem sido utilizada no campo do aprendizado de máquina com bastante sucesso em diferentes abordagens desde meados de 1985 com o trabalho de Schaffer and Grefenstette (1985), que viabilizou a implementação de algoritmos genéticos com múltiplas funções de aptidão para adaptar a estrutura da máquina de aprendizagem. O trabalho no qual o treinamento de redes neurais foi formulado como um problema de otimização multiobjetivo foi descrito em Liu and Kadirkamanathan (1995), onde duas medidas de erros de ajuste aos dados (usando norma L_2 e norma L_∞) e uma medida de complexidade (número de pesos não nulos) de uma rede de função polinomial Volterra e uma rede de função de base radial gaussiana foram minimizadas usando uma abordagem min-max, conforme a formulação destacada em (2.12):

$$\begin{aligned} F(W) &= \min_W\{\max\{f_1'(W), f_2'(W), f_3'(W)\}\} \\ f_1(W) &= \|y(W) - y^d(W)\|_2 \\ f_2(W) &= \|y(W) - y^d(W)\|_\infty \\ f_3(W) &= C \end{aligned} \tag{2.12}$$

onde C é o número de pesos não nulos, $f_1'(W), f_2'(W), f_3'(W)$ são valores normalizados de $f_1(W), f_2(W), f_3(W)$, $y(W)$ é a saída estimada, $y^d(W)$ é a saída desejada e W é a matriz de pesos da rede neural. Um algoritmo genético de objetivo único foi adotado para implementar o aprendizado e, como resultado, apenas uma solução foi alcançada.

A implementação de aprendizagem multiobjetivo por meio da abordagem por escalarização para tratar os objetivos conflitantes e a necessidade de uma abordagem que considere o *trade-off* usando o conceito de otimalidade de Pareto foi discutida em Matsuyama (1996). Os autores destacaram que métodos baseados em obtenção de soluções Pareto-ótimas eram mais apropriados.

Um importante avanço foi dado em (Kottathra and Attikiouzel, 1996), onde o treinamento de uma rede neural do tipo MLP foi formulado como um problema de otimização biobjetivo. O erro do ajuste e o número de neurônios ocultos da rede foram considerados para formulação dos objetivos. Um algoritmo *branch and bound* foi empregado para

resolver o problema multiobjetivo inteiro misto. Dada as limitações do algoritmo, a vantagem da abordagem baseada no conceito Pareto-ótimo do aprendizado de máquina não foi totalmente demonstrada no trabalho.

Um dos primeiros trabalhos a dar um tratamento mais formal para o assunto foi o de Teixeira et al. (2000), que apresentou um novo esquema para aumentar a capacidade de generalização das redes MLP. O algoritmo de treinamento multiobjetivo adotou o erro de ajuste aos dados de treinamento e a norma dos pesos como os objetivos e comparou os resultados com o algoritmo *backpropagation* padrão (Rumelhart et al., 1986) e as SVMs. Este método foi nomeado MOBJ, em referência ao termo multiobjetivo. Neste trabalho os autores usaram uma abordagem por escalarização para gerar as soluções Pareto-ótimas. A Equação 2.13 apresenta a formulação bi-objetivo do treinamento da rede MLP. Os mesmos autores apresentaram ainda trabalhos posteriores que vieram para completar a abordagem, como em de Albuquerque Teixeira et al. (2001) e a tese de doutorado de Teixeira (2001).

$$\text{Minimizar} \begin{cases} f_1(w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i, w))^2 \\ f_2(w) = ||w|| \end{cases} \tag{2.13}$$

onde w é o vetor de pesos sinápticos da rede, N é o número de dados de treinamento, y_i e $\hat{f}(x_i, w)$ são, respectivamente, a saída esperada e a saída obtida correspondente ao i-ésimo exemplo. Estas duas funções objetivo são conflitantes em uma determinada região. As soluções geradas pelo método MOBJ, no plano dos objetivos, são não-dominadas. Nesse conjunto há uma solução que mais se adequa à distribuição D dos dados. A seleção do modelo final dentre àqueles no conjunto de soluções não-dominadas é feita pelo método de validação (*holdout*) e é descrito em Teixeira (2002). O erro apresentado pelo modelo selecionado pelo método de validação, geralmente, é uma aproximação razoável do erro de predição (erro de generalização).

A formulação apresentada por Teixeira et al. (2000) foi adaptada posteriormente para uma abordagem do treinamento multiobjetivo de RBFs (*Radial Basis Function*) (Kokshenev and Braga, 2008b). Em seguida, os mesmos autores apresentaram em Kokshenev and Braga (2010) a relação entre a abordagem multiobjetivo e o princípio de minimização do risco estrutural. Além disso a seleção do modelo final usou outros critérios clássicos de seleção de modelos, como AIC e BIC, para confrontar com os resultados da seleção pelo método de validação. Também em Vieira et al. (2010) são apresentados avanços no processo de minimização do risco estrutural com uso de algoritmos multiobjetivos para controle de complexidade de uma estrutura rede chamada Perceptrons de Camadas Paralelas (PLP - *Parallel Layers Perceptron*) (Caminhas et al., 2003). Inclusive as SVMs, que também implementam de modo eficiente o princípio do risco estrutural mínimo foram estudadas sob o ponto de vista multiobjetivo em alguns trabalhos, dos quais pode-se destacar o trabalho de Suttorp and Igel (2006) e o Datta and Das (2018).

Com a crescente popularização dos algoritmos evolucionários multiobjetivo (Zitzler and Thiele, 1999; Deb, 2001), a ideia do uso desses algoritmos para os problemas de aprendizado tornou-se mais clara e prática. Em Llorà et al. (2002) foi destacado o princípio de um sistema de aprendizagem evolucionário capaz de equilibrar a acurácia e a

complexidade (referenciada como parcimônia no artigo) de modelos com boa performance geral. Em Pappa et al. (2004) os autores abordaram o uso de algoritmos evolucionários multiobjetivo na tarefa da seleção de características. Em Abbass (2003b) os autores apresentaram o estudo de algoritmos evolucionários multiobjetivo para reduzir o tempo de treinamento se comparado à técnicas baseadas em gradiente. Os trabalhos Mukhopadhyay et al. (2014); Mukhopadhyay et al. (2013) apresentaram uma visão geral das principais tarefas de *data mining* beneficiadas com o uso da abordagem evolucionária multiobjetivo. Para a análise de *clustering*, o trabalho de Maulik et al. (2011) destaca as aplicações da abordagem evolucionária multiobjetivo em *data mining* e bioinformática. Em Du et al. (2014) foi apresentada a abordagem evolucionária multiobjetivo para o problema da previsão de séries temporais. Um algoritmo de classificação baseado no princípio de dominância de Pareto é destacado em Zhang et al. (2015a) e Zhang et al. (2015b) onde foi realizada uma pré-seleção das soluções candidatas. Ainda em Kaoutar and Mohamed (2017) e Senhaji et al. (2019), foi proposto um método para geração de soluções não-dominadas de modelos neurais utilizando uma abordagem evolucionária que apresentou uma representação da fronteira de Pareto, segundo os resultados apresentados neste dois trabalhos.

Ainda há uma grande quantidade de estudos na literatura de algoritmos evolucionários multiobjetivo que possuem relação com a aprendizagem de máquina, mas não é o escopo desse trabalho.

Há um bom número de recentes trabalhos investigando os resultados da abordagem multiobjetivo de aprendizagem por reforço como foram destacados em Van Moffaert et al. (2013); Vamplew et al. (2011); Moffaert and Nowe (2014); Liu et al. (2015); Mossalam et al. (2016); Vamplew et al. (2017); Drugan et al. (2017). Em geral, as tarefas a serem aprendidas pelos agentes autônomos, para refletirem melhor o mundo real, são melhor descritas através de múltiplos objetivos.

A formulação multiobjetivo de *clustering* de dados possui alguns trabalhos que visam melhorar a efetividade da separação dos dados e, posteriormente, da validação dos *clusters* encontrados, como apontados nos trabalhos de Law et al. (2004); Handl and Knowles (2004); Handl and Knowles, 2006).

O trabalho de Gong et al. (2019) mostrou o desempenho superior de uma abordagem multiobjetivo para aprendizagem da tarefa de classificação de séries temporais uni e multidimensionais.

Os avanços expandiram em diferentes vertentes, que poderiam ser separadas em categorias de acordo com a motivação:

a melhoria da generalização na aprendizagem : onde o objetivo do processo está na capacidade dos modelos estimarem o risco estrutural mínimo, geralmente definindo duas funções objetivo a serem minimizadas: o erro e a complexidade do modelo, como destacados em Llorà and Goldberg (2003); Costa et al. (2007); Graning et al. (2006); Kokshenev and Braga (2008b);

aprimoramento da interpretabilidade na extração de regras : onde o processo de extração de regras lógicas ou fuzzy a partir de dados pode ser melhor controlado

para produzir menores erros de classificação e regras mais interpretáveis, como destacado em Ishibuchi and Nojima (2007);

geração de *ensembles* de modelos diversificados : onde a diversidade dos modelos do *ensemble* é controlada para obter um conjunto de soluções Pareto-ótimas bem distribuído, como destacado em Abbass (2003a); Smith and Jin (2014); Chen and Yao (2010).

Essas três categorias foram discutidas em Jin and Sendhoff (2008) onde as motivações para o desenvolvimento dos trabalhos são exploradas com maior profundidade. Com o crescimento dessa linha de pesquisa, foi editado um livro sobre o tema, intitulado *Multi-objective Machine Learning* (Jin, 2006), reunindo diversas linhas de pesquisa envolvendo a visão da otimização multiobjetivo em diversos aspectos da aprendizagem de máquina. No capítulo escrito por Braga et al. (2006) foi abordado um tratamento mais formal a abordagem multiobjetivo do treinamento de redes neurais, uma extensão dos resultados gerais do método MOBJ até àquele período a partir da pesquisa iniciada por Teixeira et al. (2000).

No contexto deste trabalho de tese, definido como a melhoria da generalização dos modelos de máquinas de aprendizagem, particularmente as RNAs, ficou definido que há duas etapas importantes:

- Encontrar um conjunto de soluções o mais próximo da fronteira de Pareto e, também, o melhor distribuído.
- Encontrar uma forma para a tomada de decisão no conjunto Pareto-ótimo.

Para este trabalho, dadas as duas etapas anteriormente mencionadas, a atenção ficou voltada para a etapa de tomada de decisão no conjunto Pareto-ótimo, como o da Figura 2.4, visando obter abordagens inovadoras do processo de seleção de modelos, sem o uso de reamostragem e garantindo modelos com baixo erro de predição.

Os métodos para construção de um conjunto de soluções Pareto-ótimas, além dos métodos baseados em algoritmos evolutivos (Coello et al., 2007) já mencionados neste capítulo, podem ser baseados em abordagens de otimização simples dos parâmetros da rede. Diversos métodos já foram implementados para obtenção das soluções Pareto-ótimas, como: o somatório dos pesos (MacKay, 1992) que reúne todos os objetivos em uma única função objetivo ponderada que deverá ser otimizada e a cada execução do algoritmo de otimização o termo de regularização, habitualmente referido por λ é alterado para gerar uma nova solução; o método ϵ-restrito, usado por Teixeira et al. (2000) e que define um objetivo como a função a ser otimizada e os demais objetivos como restrições. Outros métodos devem ser mencionados, como o algoritmo LASSO (*Least Absolute Shrinkage and Selection Operator*) multiobjetivo, destacado em Costa and Braga (2006) e Costa et al. (2009). O algoritmo de controle por modos deslizantes multiobjetivo (Costa et al., 2003) que é capaz de guiar a trajetória da rede MLP dentro do plano biobjetivo: erro e norma dos pesos, sendo capaz de gerar soluções não-dominadas. O algoritmo Levenberg-Marquardt adaptado para treinamento de redes MLP (Costa et al., 2007), capaz de

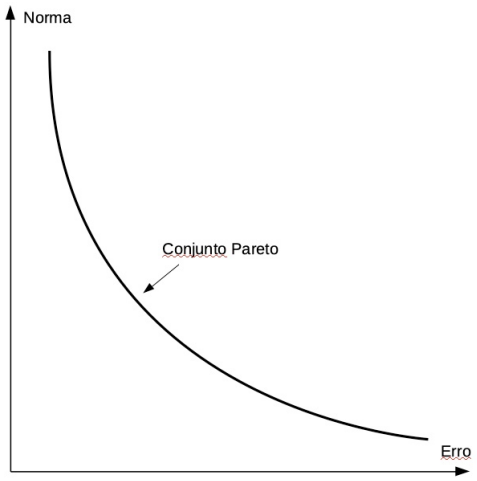


Figura 2.4: Conjunto Pareto-ótimo.

restringir o valor de norma dos pesos para um valor pré-estabelecido e encontrar a solução de erro mínimo para esse valor de norma. Em Costa and Braga (2011) discutiu-se a sofisticação de alguns dos algoritmos anteriormente mencionados e foi proposto um algoritmo acessível que decompõe o gradiente em dois componentes e ajusta os pesos da rede separadamente, realizando o treinamento multiobjetivo com o controle da norma L_2 dos pesos.

Outro aspecto discutido em abordagens multiobjetivo é a métrica adotada para controlar a complexidade do modelo. No caso das redes MLP normalmente adotou-se frequentemente a norma dos pesos, mas alguns trabalhos recentes apresentam abordagens alternativas como em Yeung et al. (2016) que descreveu em seu trabalho uma medida de complexidade baseada na sensibilidade estocástica (ST-SM) da rede. Esta ST-SM determina a expectativa das diferenças quadráticas das saídas entre as amostras de treinamento e as amostras não vistas localizadas dentro de uma vizinhança para um dado modelo Pareto-ótimo, fornecendo uma medida da flutuação das saídas de cada modelo.

Ainda na mesma linha do treinamento multiobjetivo de redes MLP, os trabalhos de Rocha et al. (2015) e Rocha (2017) apresentaram uma nova representação para os pesos, sob a forma de coordenadas esféricas, que se mostraram mais precisas na estimativa da fronteira de Pareto que a abordagem clássica usando a norma euclidiana.

A partir da aproximação do conjunto Pareto-ótimo, onde há um número limitado de soluções candidatas, a etapa seguinte resume-se no processo de escolha de uma solução (uma função $\hat{f}(x; w^*)$) pertencente a este conjunto. Para cada tipo de problema de otimização existem informações que devem ser consideradas para uma tomada de decisão. No aprendizado supervisionado é necessário que sejam elaboradas estratégias visando a escolha da solução que apresente o menor erro de generalização. A implementação de um decisor sempre é considerada uma importante etapa da otimização multiobjetivo uma vez que os algoritmos geram um conjunto de soluções candidatas e, segundo algum critério, é preciso uma escolha para finalizar com sucesso todas as etapas de um problema de

otimização (Parreiras, 2006).

A seguir, será descrito o princípio da tomada de decisão no conjunto Pareto-ótimo em problemas de aprendizado supervisionado de máquina por minimização do erro de validação. A consequência desse princípio foram as estratégias elaboradas para os problemas de classificação e de regressão.

2.4 A Regra de Decisão

O *problema de regressão* a ser abordado nesta tese é definido da seguinte forma:

Problema de Regressão: *Seja uma função geradora* $f_g : \mathbb{R} \mapsto \mathbb{R}$*:*

$$y = f_g(x) \tag{2.14}$$

em que $y \in \mathbb{R}$ *é a saída da função e* $x \in \mathbb{R}$ *é a entrada da função. Supõe-se que esteja disponível um conjunto de* N_T *medições* $[y_i]$ *da saída dessa função para um conjunto de entradas* $[x_i]$*, sendo que tais medições encontram-se corrompidas por um ruído* $[\xi_i]$*, obedecendo à relação:*

$$y_i = f_g(x_i) + \xi_i \quad ; \quad i \in \{1, \ldots, N_T\} \tag{2.15}$$

Supõe-se que o ruído $[\xi]$ *tenha variância* $\mathbb{E}(\xi^2) = \sigma$ *e média* $\mathbb{E}(\xi) = 0$*, e que não seja correlacionado nem com* x*, nem com nenhuma função determinística de* x*, ou seja:*

$$\mathbb{E}(\xi D(x)) = 0 \tag{2.16}$$

sendo $D(x)$ *uma função determinística qualquer de* x*. Seja também uma rede neural do tipo MLP cuja expressão entrada-saída é representada por:*

$$\hat{y} = f(x, w) \tag{2.17}$$

em que $\hat{y} \in \mathbb{R}$ *representa a saída da rede,* $x \in \mathbb{R}$ *representa a entrada da rede e* $w \in \mathbb{R}^m$ *representa o vetor de pesos da rede. Deseja-se determinar o vetor de pesos* w^* *tal que a função* $f(x, w^*)$ *constitua a melhor aproximação possível da função geradora dentre as funções possíveis de serem representadas por essa rede, no sentido de que tal vetor minimiza o funcional de risco* $R(w)$ *dado por:*

$$R(w) = \int_a^b (f_g(x) - f(x, w))^2 \, dx \tag{2.18}$$

em que $[a, b]$ *representa o intervalo de interesse para a aproximação.* ■

O *problema de classificação* aqui abordado é definido da seguinte forma:

Problema de Classificação: *Seja uma função indicadora* $f_g : \mathbb{R}^n \mapsto \{0, 1\}$*:*

$$y = f_g(x) \tag{2.19}$$

em que $y \in \{0, 1\}$ *é a saída da função e* $x \in \mathbb{R}^n$ *é a entrada da função. Supõe-se que esteja disponível um conjunto de* N_T *medições* $[y_i]$ *da saída dessa função para um conjunto de entradas* $[x_i]$*, sendo que tais medições podem indicar o valor correto ou o valor errado da classe de* x*, de acordo com a seguinte distribuição de probabilidades:*

$$\begin{cases} P(y_i = f_g(x_i)) = p \\ \\ P(y_i = \bar{f}_g(x_i)) = 1 - p \end{cases} \tag{2.20}$$

onde $\bar{f}_g(x_i)$ *significa o complemento de* $f_g(x_i)$*, e* $p \in [0, 1]$ *é a probabilidade de acerto. Supõe-se que essa distribuição de probabilidades não tenha dependência nem com* x *nem com* y*. Seja uma rede neural do tipo MLP cuja expressão entrada-saída é representada por:*

$$\hat{y} = f(x, w) \tag{2.21}$$

em que $\hat{y} \in \{0, 1\}$ *representa a saída da rede,* $x \in \mathbb{R}^n$ *representa a entrada da rede e* $w \in \mathbb{R}^m$ *representa o vetor de pesos da rede. Deseja-se determinar o vetor de pesos* w^* *tal que a função* $f(x, w^*)$ *constitua a melhor aproximação possível da função geradora dentre as funções possíveis de serem representadas por essa rede, no sentido de que tal vetor minimiza o funcional de risco* $R(w)$ *dado por:*

$$R(w) = \int_{\mathcal{R}} (f_g(x) - f(x, w))^2 \, dV \tag{2.22}$$

em que $\mathcal{R}$ *representa a região de interesse para a aproximação.* ■

É claro que não se tem acesso direto ao funcional $R(w)$ em nenhum dos casos, de forma que não é possível proceder diretamente à sua minimização. A estratégia aqui adotada é constituída de duas etapas:

1. Primeiro, realiza-se um procedimento de *otimização multiobjetivo*, utilizando um conjunto contendo N_T pares (y_i, x_i), com a minimização dos seguintes funcionais:

$$\begin{cases} f_1(w) = \sum_{i=1}^{N_T} [y_i - f(x_i, w)]^2 \\ \\ f_2(w) = \|w\|_2^2 = \sum_{i=1}^{m} w_i^2 \end{cases} \tag{2.23}$$

   Esse procedimento irá gerar um conjunto $\mathcal{W}^*$ contendo n_p vetores que representam amostras do conjunto Pareto-ótimo desse problema multiobjetivo.

2. A seguir, realiza-se uma busca sobre o conjunto discreto $\mathcal{W}^*$ para determinar o vetor de pesos $w^* \in \mathcal{W}^*$ que representa a melhor escolha possível de uma função

$f(x, w)$ para aproximar $f_g(x)$, dada a informação disponível, representada pelos N_T pares (y_i, x_i).

O trabalho desenvolvido nesta tese se concentra na busca de procedimentos adequados para a realização do segundo passo acima, no qual se escolhe w^* no conjunto $\mathcal{W}^*$. Deve-se notar que um dos objetivos definidos para esta tese é o de evitar que seja necessário reservar um subconjunto dos dados disponíveis para a execução desta etapa, assim permitindo que a primeira etapa utilize todos os N_T pares (y_i, x_i) disponíveis, assim aumentando a acurácia do passo 1, em relação a métodos que requerem a separação de um subconjunto dos dados para esta etapa de escolha.

Neste processo de decisão, alguns trabalhos apresentaram propostas de resultados promissores. O trabalho de Teixeira (2002) discutiu o princípio do decisor original do método MOBJ (Teixeira et al., 2000) garantindo que o método baseado no erro de validação apresentava uma estimativa razoável erro de generalização. Em Teixeira et al. (2007) foi apresentada uma abordagem de treinamento e, ao mesmo tempo de seleção de modelo, que adotou um critério baseado na busca *golden section* para direcionar o treinamento na geração de soluções eficientes até obter a solução final de boa capacidade de generalização. Em Medeiros (2007) foi apresentado um resultado inicial do processo de decisão para problemas de regressão, baseado na avaliação da autocorrelação do sinal resídual dos modelos Pareto-ótimo, inspirado no trabalho de Barroso et al. (2007) que discutiu o princípio de mínima autocorrelação na estimação de parâmetros de um modelo polinomial NARX . Em Medeiros et al. (2009) resultados iniciais de uma proposta baseada em incorporação de conhecimento *a priori* foi discutida como decisor. No trabalho de Kokshenev and Braga (2010) apresentou-se brevemente a possibilidade de uso dos critérios clássicos de seleção de modelos (AIC e BIC) na abordagem multiobjetivo para RBFs. No trabalho de Torres et al. (2012) implementou-se um decisor baseado na geometria da margem de separação entre as classes para realizar a seleção do modelo Pareto-ótimo. Em Medeiros et al. (2017) mostrou-se a validade do uso de informações *a priori* para modelagem de um decisor na seleção de modelos Pareto-ótimos em problemas de classificação.

Ainda há um método heurístico para escolha do parâmetro de regularização em problemas mal-postos baseado na região de máxima curvatura da curva-L (Hansen, 1992; Hanke, 1996; Hansen, 2001; Kindermann and Raik, 2019), que será discutido em um capítulo próprio mais à frente.

2.5 Comentários Finais

Neste capítulo discutiu-se brevemente o problema geral da seleção de modelos dentro da abordagem multiobjetivo de aprendizagem de máquina e como ela está relacionada ao princípio de minimização do risco estrutural. O interesse em descrever o princípio geral da tomada de decisão *à posteriori* em problemas de aprendizado multiobjetivo a partir da abordagem de seleção por erro de validação gerou a base para a formulação de novas estruturas de decisão. A hipótese fundamental usada para a construção do problema geral

de decisão baseia-se na condição da existência de um ruído nos atributos-meta (classes) do conjunto de dados. Nos capítulos seguintes, serão destacadas as duas abordagens de decisão capazes de escolher uma solução de menor erro de generalização e que constituem nas contribuições fundamentais deste trabalho.

Capítulo 3

A Curva L e a Curva de Pareto

É atribuído a Oleg Mikailivitch Alifanov, proeminente pesquisador russo na área de problemas inversos, a afirmação "a solução de um problema inverso consiste em determinar causas baseado na observação dos seus efeitos".

A apresentação do que consiste um problema inverso, deve-se ao fato da Teoria de Regularização, baseada no método conhecido como *Curva L*, utilizar uma abordagem bastante similar ao método MOBJ.

3.1 Problemas Inversos

É creditado ao astrofísico georgiano Viktor Amazaspovich Ambartsumian haver cunhado a expressão Problema Inverso (PI). Uma definição bastante abrangente, porém, é apresentada em Engl et al. (1996): "Resolver um problema inverso é determinar causas desconhecidas a partir de efeitos desejados ou observados". Note-se que a área de projeto ótimo ou projeto inverso (*inverse design*) também está incluída nesta definição.

De acordo com a Figura 3.1 é possível mostrar a relação entre um problema direto e um inverso. As *causas*, num modelo matemático, são as condições iniciais e de contorno, termo de fontes/sumidouro e propriedades do sistema. Os *efeitos* são as propriedades calculadas a partir de um modelo direto, como o campo de temperatura, corrente elétrica etc.

As *causas* (ou natureza) podem ser usadas para determinação da classe do PI, apesar de outras classificações serem possíveis.

Natureza Matemática: explícito ou implícito;

Natureza Estatística: determinística ou estocástica;

Natureza da Propriedade: condição inicial, condição de contorno, termo de fonte/sumidouro, propriedades do sistema;

Natureza da Solução: estimação de parâmetros, estimação de função.

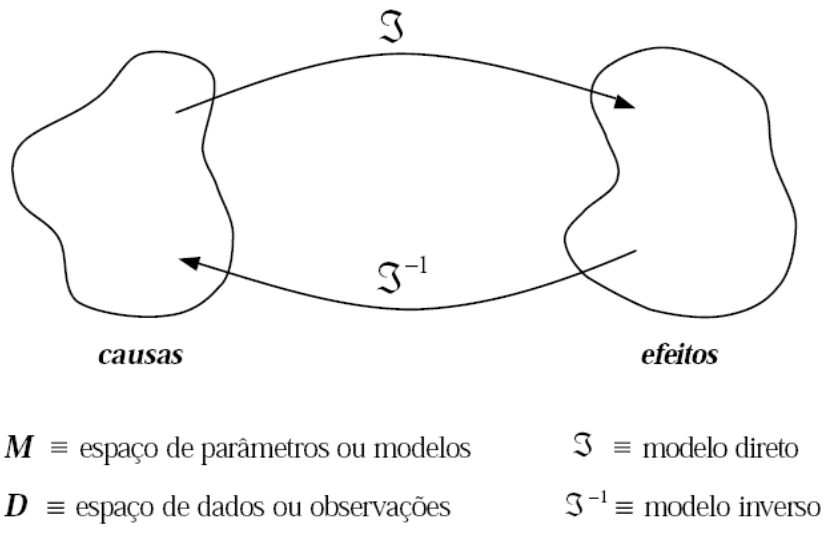


Figura 3.1: Representação de um Problema Direto e Inverso.

Matematicamente, os problemas inversos frequentemente pertencem à classe dos problemas mal-postos. No início do século XX, o matemático francês Jacques Hadamard definiu um problema bem-posto como sendo aquele que cumpre as seguintes condições:

- (i) existência de uma solução;
- (ii) a solução é única;
- (iii) a solução têm dependência contínua (suave) com os dados de entrada.

Portanto, um problema é chamado mal-posto se alguma das condições descritas por Hadamard não é satisfeita. Problemas discretos e finitos são chamados mal condicionados, se a condição (iii) não é satisfeita. Em um problema inverso basta que uma das condições de Hadamard não seja satisfeita para termos um problema mal-posto.

É possível elaborar um paralelo entre um problema de aprendizado e um problema inverso. Em especial, problemas de aprendizado supervisionado de máquina podem ser classificados como problemas inversos (Kurková, 2005; Vito et al., 2005). Tais problemas são estudados neste trabalho

Dentre um conjunto de métodos para solução dos PIs, o método de regularização é bastante utilizado para obter-se um modelo bem-posto por meio da utilização de algum conhecimento prévio do problema.

3.1.1 Teoria de Regularização

A Teoria de Regularização surgiu como um método proposto por Tikhonov and Arsenin (1977), para resolver problemas mal-postos. A ideia básica da regularização é *estabilizar* a solução por meio de algum funcional não-negativo que incorpore informação prévia sobre a solução. A forma mais comum de informação prévia envolve a suposição de que a função do mapeamento entrada-saída seja *suave*, no sentido de que entradas similares correspondam a saídas similares. A busca por um mapeamento entrada-saída mais *suave* é uma informação adicional, que transforma um problema mal-posto em um problema bem-posto, como na Figura 3.2.

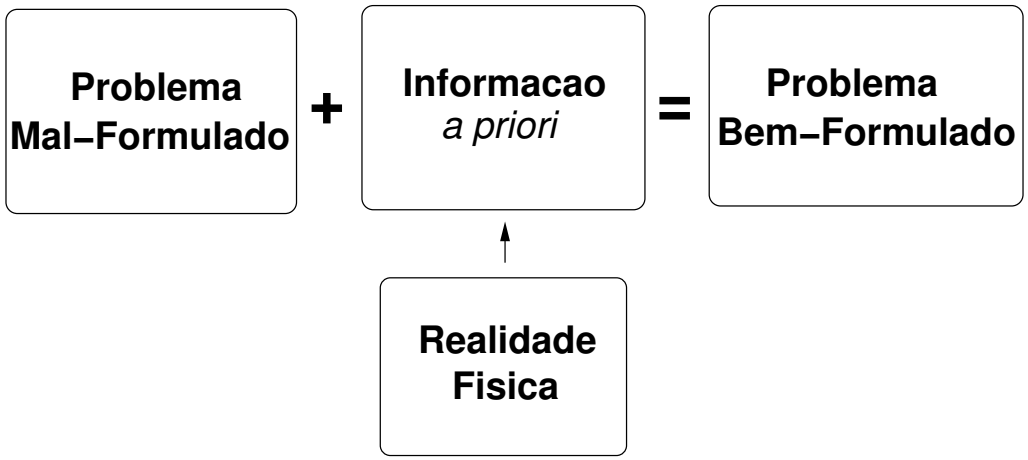


Figura 3.2: Ideia Básica da Teoria da Regularização

Considerando uma amostra de treinamento como:

$$\begin{aligned} \text{Entrada:} \quad & x_i \in \Re^{m_0}, \;\; i = 1, 2, ..., N \\ \text{Saída Desejada:} \quad & d_i \in \Re^1, \;\; i = 1, 2, ..., N \end{aligned} \tag{3.1}$$

Nota-se que apesar da suposição de que a saída seja uni-dimensional, a aplicabilidade geral da teoria da regularização não é limitada a esta suposição. Basicamente, a teoria de regularização envolve dois termos:

1. *Termo de Erro Padrão.* Representado por $E_D(w)$, mede o erro padrão entre a resposta desejada d_i e a resposta real obtida y_i para os padrões de treinamento $i = 1, 2, ..., N$. Define-se

$$\begin{aligned} E_D(w) &= \frac{1}{2}\sum_{i=1}^{N}(d_i - y_i)^2 \\ &= \frac{1}{2}\sum_{i=1}^{N}[d_i - f(x_i)]^2 \end{aligned} \tag{3.2}$$

onde $f(x_i)$ é uma função aproximada do problema.

2. *Termo de Regularização*. Este termo, representado por $E_S(w)$ depende das propriedades "geométricas" da função aproximada $f(x)$. Especificamente, pode-se escrever

$$E_S(w) = \Omega(w) \tag{3.3}$$

onde Ω é referenciado como um funcional *estabilizador* porque ele estabiliza a solução para o problema de regularização, fazendo-a suave e, desta maneira, satisfazendo à propriedade de continuidade. A suavidade implica na continuidade, porém o contrário nem sempre é verdade.

Desta forma a função a ser minimizada na teoria de regularização é

$$\begin{aligned} J(w) &= E_D(w) + \lambda E_S(w) \\ &= \frac{1}{2}\sum_{i=1}^{N}[d_i - f(x_i)]^2 + \lambda\Omega(w) \end{aligned} \tag{3.4}$$

onde λ é um número real positivo denominado de *Parâmetro de Regularização* e $E_S(w)$ é denominado o *Termo de Regularização*. O minimizador para o funcional $J(w)$, a solução para o problema de regularização, é representado por $f_\lambda(x)$. De acordo com essa formulação, o *Parâmetro de Regularização* (λ) pode ser considerado como um indicador da suficiência do conjunto de dados fornecido como exemplo que especifica a solução $f_\lambda(x)$. No caso particular em que $\lambda = 0$, o problema é irrestrito, com a solução $f_\lambda(x)$ sendo totalmente determinada pelos padrões entrada-saída. No outro caso limite, $\lambda \to \infty$, implica que a restrição de suavidade imposta pelo operador $\Omega(w)$ é por si só suficiente para especificar a solução $f_\lambda(x)$, indicando que os exemplos não são confiáveis. Na prática, o *Parâmetro de Regularização* λ assume um valor entre esses dois extremos $(0, \infty)$ de forma que haja um equilíbrio entre a amostra de dados e a informação prévia. Desta forma, o *Termo de Regularização* representa uma forma de *punição de complexidade*, cuja influência sobre a solução final é controlada pelo Parâmetro de Regularização (λ).

Assim, a obtenção de um método de regularização consiste em:

1. Achar o funcional estabilizador $\Omega(w)$.

2. Escolher o parâmetro de regularização λ, preferencialmente levando-se em conta o ruído.

Há duas abordagens de funcionais estabilizadores que serão apresentados a seguir.

Regularização de Tikhonov

Em Tikhonov and Arsenin (1977) foi apresentada uma técnica de regularização definida a partir do seguinte funcional estabilizador:

$$\Omega(w) = \alpha_0||w||_2^2 + \alpha_1||w^{(1)}||_2^2 + \cdots + \alpha_p||w^{(p)}||_2^2, \tag{3.5}$$

onde $w^{(i)}$ denota a i-ésima derivada de w em relação a t para $i = 0, \cdots, p$ e os parâmetros de regularização $\alpha_p \geq 0$ são chamados estabilizadores de ordem p de Tikhonov. Por exemplo, se $\alpha_0 = 1$ e $\alpha_i = 0$ para $i = 1, \cdots, p$ então o método é chamado de *regularização de Tikhonov de ordem zero*, onde o o funcional $J(w)$ apresenta a forma:

$$J(w) = \frac{1}{2}\sum_{i=1}^{N}[d_i - f(x_i)]^2 + \lambda\Omega(w) \tag{3.6}$$

$$= \frac{1}{2}\sum_{i=1}^{N}[d_i - f(x_i)]^2 + \lambda||w||_2^2 \tag{3.7}$$

Veja que quando $\lambda \to 0$ há predominância do termo de ajuste aos dados, ou seja, busca-se um ajuste perfeito entre a saída da função aproximada e a saída dos dados. Contudo quando o $\lambda << 1$ o termo $\Omega(w)$ pode crescer bastante, na medida que o produto $\lambda\Omega(w)$ não representa um valor significativo na minimização do funcional definido em 3.6 e portanto, a magnitude $||w||_2$ da solução pode atingir valores elevados, podendo apresentar comportamento instável em relação aos dados. Mas, em contra-partida, para um λ grande as magnitudes de w diminuem e tendem a zero no limite quando $\lambda \to \infty$, ignorando então quase toda informação disponível nos dados.

Portanto, a regularização de Tikhonov é um procedimento que modifica a abordagem de minimização por mínimos quadrados através da adição de operadores regularizadores (suavizadores) que reduzem a influência do ruído sobre os resultados da aproximação. Do ponto de vista matemático isto consiste em construir um novo operador que pode gerar resultados estáveis no sentido de continuidade em relação aos dados.

O problema de seleção do parâmetro de regularização é um dos principais problemas para a abordagem através do funcional de Tikhonov. A escolha adequada do parâmetro de regularização permite obter-se soluções com flutuações reduzidas e estáveis. Dentre alguns métodos conhecidos para seleção parâmetro de regularização, está a *Curva L*.

3.1.2 A Curva L

A *Curva L* é um gráfico em escala logarítmica da norma das soluções regularizadas *versus* a norma do resíduo correspondente para um dado modelo em particular. É uma ferramenta gráfica para exibição do dilema entre a complexidade da solução regularizada $\|L\mathbf{x}_\lambda\|$ e o ajuste desse modelo aos dados $\|A\mathbf{x}_\lambda - b\|$, com vários parâmetros de regularização (λ). Deste modo, é possível visualizar o compromisso entre a minimização destas duas quantidades.

O uso prático de tal gráfico foi sugerido pela primeira vez por Lawson and Hanson (1974). Gráficos similares também surgiram em Miller (1970) e, apareceram também em Hansen (1992, 1990, 2001) e ainda em Hansen et al. (2004). É fato que vários pesquisadores têm utilizado esta forma gráfica para tratar problemas inversos mal-postos. Com a *Curva-L* é possível estimar o valor "ótimo" do parâmetro de regularização, tal como outros métodos, como a GCV (*Generalized Cross Validation*) (Golub et al., 1979) e a discrepância de Morozov (Ramlau, 2001).

Para problemas com parâmetros de regularização discretos, a *Curva L* discreta consiste em um gráfico com um conjunto de pontos $(\rho, \eta) = (\|A\mathbf{x}_\lambda - b\|_2^2, \ \|\mathbf{x}_\lambda\|_2^2)$ que são representados em escala logarítmica por $\widehat{\rho} = log\ \rho$ e $\widehat{\eta} = log\ \eta$. O conjunto de pontos

$$(\widehat{\rho}_k, \quad \widehat{\eta}_k) = (log\|A\mathbf{x}_{\lambda_k} - b\|_2, \ \ log\|\mathbf{x}_{\lambda_k}\|_2), \quad k = 1, 2, \ldots, n. \tag{3.8}$$

onde n é o número de soluções regularizadas. Em muitos problemas, de diferentes aplicações, o objetivo da regularização com o critério da *Curva L* é encontrar uma curva, contínua ou discreta (Figura 3.3), em forma de "L", aproximadamente. Com esta técnica, a solução com o parâmetro de regularização mais adequado é a solução correspondente ao ponto mais próximo ao "corner" da curva em forma de "L".

Para *Curvas L* contínuas, foi sugerido por Edelman (1988) definir que o corner é o ponto pertencente à *Curva L*, de maior curvatura. Ou seja, o ponto que defina o menor ângulo formado pela curva contínua. Já para as *Curvas L* discretas é menos evidente a forma operacional para definição do corner adequado. A Figura 3.3 mostra o valor para o parâmetro de regularização λ obtido por meio do método da *Curva L*.

Neste ponto, a *Curva L* apresenta uma relação de semelhança com o método MOBJ, no qual ambas fazem uma representação da complexidade das soluções versus o erro de ajuste para tratar o dilema entre o ajuste do modelos aos dados e a complexidade da solução.

O Critério da *Curva L* para Determinação do Parâmetro de Regularização

O fato de que a *Curva L* para problemas com dados ruidosos apresenta um corner mais ou menos distiguível, permite uma estratégia para escolha do parâmetro de regularização λ, sabendo que, um dado ponto $(\widehat{\rho}, \widehat{\eta}) = (log\|A\mathbf{x}_\lambda - b\|_2, log\|\mathbf{x}_\lambda\|_2)$ sobre a *Curva L* está neste corner (Hansen, 1992). A ideia por trás desta escolha é que o corner separa as linhas verticais, onde a solução é dominada pela regularização excessiva, e horizontais, onde a solução é dominada pelo ruído. Então o *critério da Curva L* para determinação

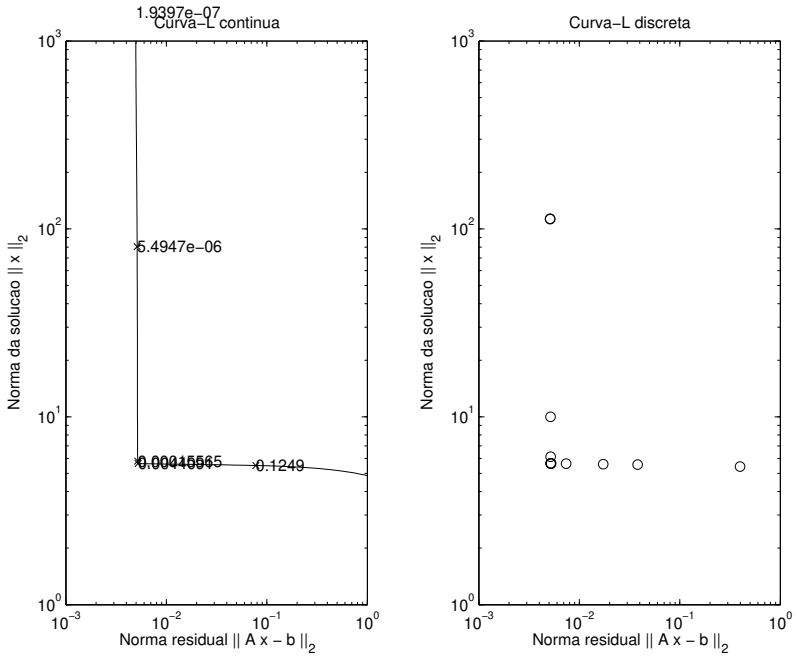


Figura 3.3: *Curvas-L* para problemas contínuos e discretos.

do parâmetro de regularização é um dos poucos métodos que envolvem, ambos, norma residual (erro) $\|A\mathbf{x}_\lambda - b\|_2$ e norma da solução $\|\mathbf{x}_\lambda\|_2$.

Para fornecer a definição estritamente matemática do "corner" da *Curva L*, utiliza-se o ponto de máxima curvatura na *Curva L*. Sabendo que $\widehat{\rho}$ e $\widehat{\eta}$ são funções de λ, e dado que $\widehat{\rho}'$, $\widehat{\eta}'$, $\widehat{\rho}''$ e $\widehat{\eta}''$ são as derivadas primeira e segunda de $\widehat{\rho}$ e $\widehat{\eta}$ em relação à λ. Então, a curvatura k da *Curva L* $(\widehat{\rho}, \widehat{\eta})$, como uma função de λ, e dado por Hanke (1996):

$$k = \frac{\widehat{\rho}'\widehat{\eta}'' - \widehat{\rho}''\widehat{\eta}'}{((\widehat{\rho}')^2 + (\widehat{\rho}')^2)^{3/2}}. \tag{3.9}$$

Assim, é simples usar um procedimento de otimização uni-dimensional para calcular o máximo k, dado pela Equação 3.9. Existem outros métodos para determinação do "corner" da *Curva L* apresentados por Hansen and O'Leary (1993).

A Figura 3.4 mostra uma *Curva L*, onde o corner é facilmente distinguível e a curvatura associada da *Curva L* como uma função de λ. O pico no k-ésimo valor de λ da *Curva L* corresponde ao "corner" sobre a *Curva L*.

Comparações experimentais do critério da *Curva L* com outros métodos para cálculo de λ, principalmente com a *Generalized Cross Validation - (GCV)* foram apresentados por Hansen and O'Leary (1993). A conclusão tirada a partir desses experimentos foi que o critério baseado na *Curva L* mostrou-se mais robusto em problemas com um ruído aditivo enquanto o método GCV ocasionamente falhava. Porém o critério da *Curva L*

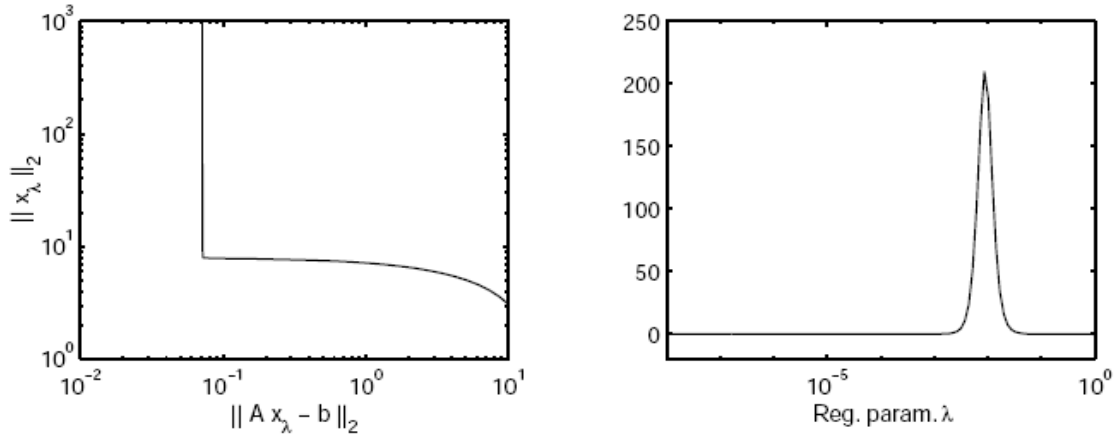


Figura 3.4: Uma *Curva L* típica e um gráfico da curvatura k em função do parâmetro de regularização λ.

apresenta uma tendência em produzir uma solução sub-ajustada aos dados, ou seja, um λ muito grande. Experimentos com ruídos correlacionados, ainda no trabalho de Hansen and O'Leary (1993), mostraram que o critério da *curva-l* é superior ao método GCV, que apresentou resultados fortemente sobre-ajustados, ou seja, λ muito pequeno.

Limitações do Critério Curva L

Como todos métodos práticos, este possui vantagens e desvantagens. As vantagens da *Curva L* são a robustez e a habilidade para tratar com problemas corrompidos por um ruído correlacionado. Porém, duas desvantagens ou limitações destacadas em detalhes em (Hansen and O'Leary, 1993; Vogel, 1996) serão brevemente discutidas. A compreensão destas limitações é fundamental para o uso apropriado do critério da *Curva L* e também para futuras melhorias no método.

A primeira limitação é a reconstrução de uma solução exata muito suave, tornando o parâmetro de regularização algumas ordens de grandeza menores que o parâmetro ótimo.

A segunda limitação do critério da *Curva L* está relacionada a seu comportamento assintótico enquanto o tamanho do espaço de variáveis n do problema aumenta. Como indicado por Vogel (1996), o parâmetro de regularização obtido pelo critério da *Curva L* pode não comportar-se consistentemente enquanto a dimensão n aumenta.

Em relação ao aprendizado multiobjetivo, é possível analisar as seguintes características da *Curva L*: a limitação desse critério em mapear soluções regularizadas na região não-convexa do conjunto Pareto-ótimo, quando esta existir.

Em uma análise cuidadosa, a *Curva L* e Curva de Pareto são representações semelhantes do mesmo problema: a determinação do modelo de melhor ajuste aos dados, equilibrando dois termos: a complexidade do ajuste (norma dos parâmetros livres) e o erro (resíduo) entre o modelo aproximado e os dados. Porém, originalmente a escala

da curva de Pareto não é logarítmica, o que visualmente não tornava essa analogia tão evidente. Porém, alterando a escala da curva de Pareto para uma escala logarítmica, é possível perceber a semelhança entre os métodos para construção de diversas soluções e, posteriormente, a escolha da melhor solução.

3.1.3 A Curva de Validação

A proposta inicial de decisão implementada no método MOBJ foi baseada em um processo de amostragem dos dados do problema de aprendizado. Este processo permitia ao decisor obter uma curva denominada curva de validação que, ao contrário da curva de Pareto, obtida com os dados de treinamento, apresentava um ponto de mínimo. A decisão implementada escolhe a solução correspondente a esse ponto de mínimo erro obtido na curva discreta de validação. Conforme destacado por Teixeira (2002), o decisor baseado no princípio do mínimo erro de validação é capaz de assegurar a escolha de uma solução com alta capacidade de generalização, dentre aquelas que compõem o conjunto Pareto-ótimo.

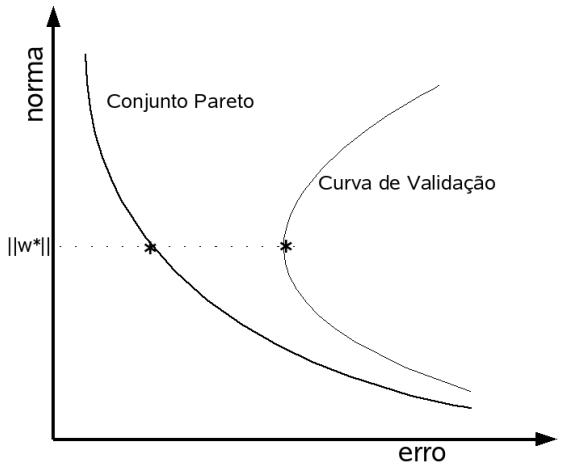


Figura 3.5: Curva de Pareto e a Curva de Validação.

A curva de validação é obtida por meio de uma avaliação dos modelos gerados pelo treinamento multiobjetivo com um novo conjunto de dados. Como o conjunto Pareto-ótimo é composto por modelos sub-ajustados até os super-ajustados, espera-se que o modelo de melhor ajuste, seja o que apresentar o menor erro para o conjunto de dados de validação. Portanto, os modelos Pareto-ótimos com a complexidade inferior e os que apresentam complexidade superior ao necessário, terão um erro de validação elevado em relação ao modelo de melhor ajuste. Desse modo, o decisor baseado em erro de validação busca a região de mínimo erro na curva de validação. Essa região de mínimo, corresponde ao modelo de melhor capacidade de generalização dentre os modelos com conjunto Pareto-ótimo.

A curva de validação, conforme é apresentada na Figura 3.5 apresenta uma região onde está localizado um ponto de mínimo em problemas de aprendizado com ou sem a presença de um ruído nos dados. A presença de uma região de mínimo nesta curva indica a região que contém a solução de melhor capacidade de generalização, segundo os dados utilizados para validar cada solução Pareto-ótima. O ponto de mínimo erro de validação não irá mudar em relação à complexidade da solução, medida pela norma euclidiana.

A estratégia de decisão baseada em um conjunto de dados de validação deve ser consistente, ou seja, mesmo que os dados de validação estejam corrompidos por um ruído de natureza aleatória, a localização da solução de mínimo erro de validação não deve sofrer grandes variações em sua medida de complexidade. A Figura 3.6 mostra o deslocamento esperado da curva de validação à medida que o ruído não correlacionado apresenta um aumento na medida de variância.

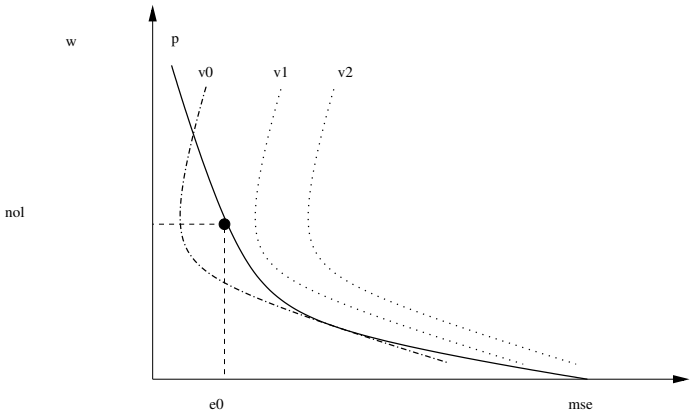


Figura 3.6: Deslocamento da região de mínimo erro na curva de validação em problemas com ruído não correlacionado.

Espera-se que a medida do mínimo erro de validação esteja relacionada com a medida da variância do ruído. Como foi argumentado no capítulo 2, o erro encontrado na avaliação do modelo, como a rede MLP por exemplo, deve estar relacionado com a medida da variância do ruído. Idealmente, a solução deste decisor não deve ser afetada. A relação entre ruído dos dados e erro de validação das soluções deverá ficar presente na medida do erro, que para cada solução Pareto-ótima estima-se que seja a variância do ruído. O problema é que essa característica do ruído é desconhecida na maioria das situações reais.

Na Figura 3.7, é apresentado um deslocamento do mínimo da curva de validação a partir do instante em que o problema de aprendizagem possui um ruído correlacionado com a função geradora dos dados, o que acarreta na escolha de um modelo que deixa de ser o melhor para o problema. Neste caso, o ruído não é uma variável totalmente aleatória, o que dificulta o processo de aprendizagem. Para três diferentes níveis de ruído c_1, c_2 e c_3, cada um com um nível de correlação maior em relação à função geradora, espera-se que o decisor por mínimo erro de validação escolha o modelo com o valor de norma cada vez maior, uma vez que este cenário é mais problemático para este decisor.

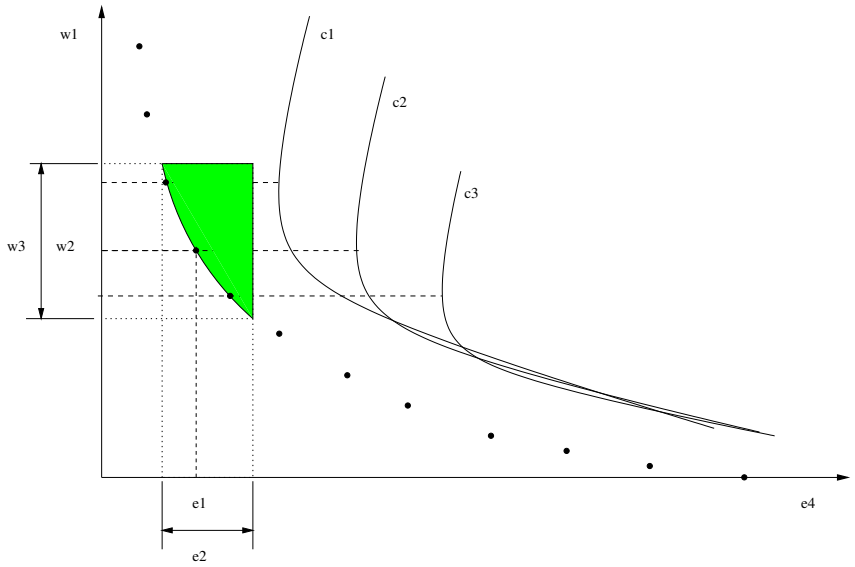


Figura 3.7: Deslocamento da região de mínimo erro na curva de validação em problemas com ruído correlacionado.

A região preenchida na Figura 3.7 mostra uma condição em que as três soluções obtidas podem ser equivalentes, uma vez que o nível de correlação do ruído permite que seja relaxada a condição de uma única solução escolhida e define uma sub-região do conjunto Pareto-ótimo, onde os modelos treinados pelo método MOBJ são considerados equivalentes dentro de um intervalo de variação.

Apesar da viabilidade do decisor de mínimo erro de validação, sua escolha no conjunto de soluções Pareto-ótimas pode sofrer de um viés originário de sua definição. Este decisor, baseado nos dados de validação pode ter a tendência de numa região de equivalência entre as soluções, escolher soluções com complexidade um pouco mais elevada, ou seja, que apresentam um leve efeito de *overfitting*.

$$e_V = \frac{1}{N_V} \sum_{i=1}^{N_V} [(f(x_{V_i}) + \xi_i) - \hat{f}(x_{V_i}; w)]^2 \tag{3.10}$$

A questão central neste decisor é que sua escolha visa atender a um critério que possui um informação de ruído, e assim mesmo busca-se pela solução Pareto-ótima que minimize o critério da Equação (3.10), com N_V dados de validação. A qualidade da amostra usada para o procedimento de validação sempre será um fator determinante do sucesso dessa estratégia de decisão, seja pela quantidade de dados ou pela distribuição.

Como destacado no capítulo 2, a expectativa do processo de decisão sob a perspectiva do aprendizado como um problema de regressão tende a gerar um resíduo que tenha a magnitude da variância do ruído dos dados à medida que a aproximação é perfeita. Como o decisor é orientado pela minimização do e_V, a busca pela solução de mínimo erro

de validação pode escolher a solução Pareto-ótima com o resíduo de mínima variância. Destaco aqui que o resíduo corresponde à diferença entre saída esperada dos dados e saída estimada pela solução organizadas temporalmente quando tratamos problemas de regressão ou aproximação de função.

$$e_V = \frac{1}{N_V} \sum_{i=1}^{N_V} (e_i + \xi_i)^2 \tag{3.11}$$

onde e_i representa a estimativa de erro entre a função aproximada e a função real $e = f(x_{V_i}) - \hat{f}(x_{V_i}; w)$ e ξ representa um ruído.

Em um caso particular cujo $e_i \approx 0$, o erro será $e_V \approx \sigma^2$. Neste caso, nenhuma outra solução Pareto-ótima apresentará um e_V inferior e assim, a melhor solução será determinada pelo mínimo e_V, consequentemente um sinal com a característica de mínima variância.

Porém, em outros casos em que $e_i \neq 0$, o e_V apresentará a informação do ruído e do erro acumuladas sem identificar precisamente a contribuição de cada termo. Neste cenário, o sinal do resíduo com a mínima variância pode não corresponder a solução que melhor represente a função real. Isto porque na prática a solução do decisor por mínimo erro de validação, que também indica mínima variância do resíduo, pode vir a escolher uma solução super-ajustada aos dados devido ao viés de sua abordagem em escolher a solução de mínima variância.

As simulações computacionais usadas neste capítulo apresentam alguns resultados simples para uma comparação entre o decisor por erro de validação do método MOBJ e a decisão pelo critério da *Curva-l*.

3.1.4 Simulações Computacionais

Foram realizados alguns experimentos numéricos, onde o método MOBJ para redes MLP gerou um conjunto de soluções Pareto-ótimas. Para análise, o conjunto Pareto sofre uma modificação em sua escala para induzir a curva em formato L. A representação da fronteira Pareto-ótimo em escala logarítmica, apresenta uma aproximação do formato "L". A região de corner desta aproximação corresponde à solução indicada pela *Curva L* e, em problemas cuja aproximação corresponde ao formato desejado, a solução escolhida pelo decisor do método multiobjetivo também corresponde à mesma região.

Portanto, uma descrição formal do relacionamento existente entre a *Curva L* e a Curva de Pareto mostrará a generalidade da abordagem multiobjetivo no aprendizado supervisionado. Os métodos de regularização são capazes de gerarem soluções Pareto-ótimas, porém especialmente no caso de a fronteira Pareto-ótima apresentar regiões não-convexas, o método da *Curva L* e outros métodos de determinação do parâmetro de regularização não são capazes de mapear essa região. Acontece que em alguns problemas reais, segundo a abordagem MOBJ, a solução encontra-se na região não-convexa da curva de Pareto. Os experimentos numéricos dessa seção mostram os resultados dessa comparação usando o decisor por mínimo erro de validação.

A Figura 3.8 mostra a *Curva L* e a aproximação (linha tracejada) da função geradora

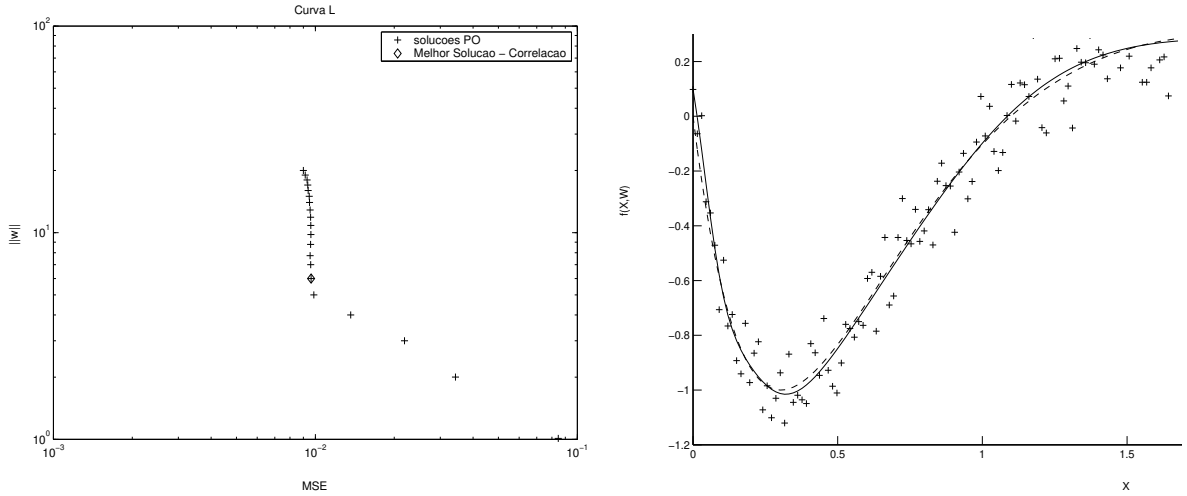


Figura 3.8: A *Curva L* (Pareto em escala logarítmica) e a melhor aproximação obtida por uma estratégia de decisão.

(linha contínua) a partir do conjunto de dados com ruído (+) disponibilizado para o aprendizado. Para outro problema de aprendizado, a Figura 3.9 mostra a *Curva L* e a aproximação (linha tracejada) da função geradora (linha contínua) a partir de um outro conjunto de dados com ruído (+) de um problema diferente.

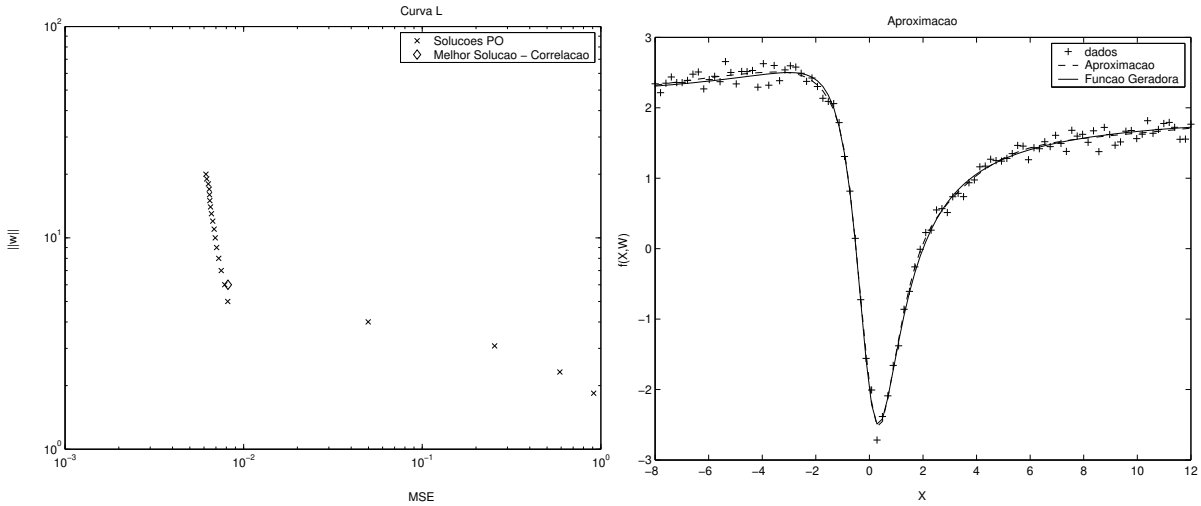


Figura 3.9: A *Curva L* (Pareto em escala logaritmica) e a melhor aproximação obtida por uma estratégia de decisão.

Nos dois problemas teste utilizados para exibir a Curva Pareto obtida em uma *Curva L*, as propriedades desejadas para uma boa decisão foram apresentadas e foi possível perceber que a melhor solução MOBJ encontra-se próximo da região do corner da *Curva*

L para estes problemas.

Classificação de Padrões - Distribuições Gaussianas

O seguinte problema de classificação de padrões (Figura 3.10) apresenta duas classes que possuem padrões amostrados aleatoriamente de duas distribuições normais bivariadas em $\Re^2$. As distribuições são dadas pelas Equações 3.12 e 3.13. Para a primeira classe, as médias em relação a cada variável são $\mu_{x1}^{c1} = \mu_{x2}^{c1} = 2$. Para a segunda classe, as médias são $\mu_{x1}^{c2} = \mu_{x2}^{c2} = 4$. Essas duas classes apresentam variâncias iguais, ou seja, $\sigma_1^2 = \sigma_2^2 = 1.5^2$. Essa variância gerou uma sobreposição dos dados. Para a classe (o) a resposta da rede deve ser $d = -1$ e para a classe (+), a resposta da rede deve ser $d = +1$.

$$p_1(x_1, x_2) = \frac{1}{2\pi\sigma_1^2} \exp\left[-\frac{1}{2}\left\{\frac{(x_1 - \mu_{x1}^{c1})^2}{\sigma_1^2} + \frac{(x_2 - \mu_{x2}^{c1})^2}{\sigma_1^2}\right\}\right] \tag{3.12}$$

$$p_2(x_1, x_2) = \frac{1}{2\pi\sigma_2^2} \exp\left[-\frac{1}{2}\left\{\frac{(x_1 - \mu_{x1}^{c2})^2}{\sigma_2^2} + \frac{(x_2 - \mu_{x2}^{c2})^2}{\sigma_2^2}\right\}\right] \tag{3.13}$$

Para o processo de treinamento, a estrutura da rede foi definida com: 2 unidades de entrada, 30 neurônios na camada oculta e 1 unidade de saída. A função de transferência (ativação) utilizada nas camadas ocultas e de saída foi a tangente hiperbólica. O método MOBJ gerou 30 soluções Pareto-ótimas com valor de norma variando em 0.5 a cada solução gerada. A Figura 3.10 ilustra a distribuição das classes do problema.

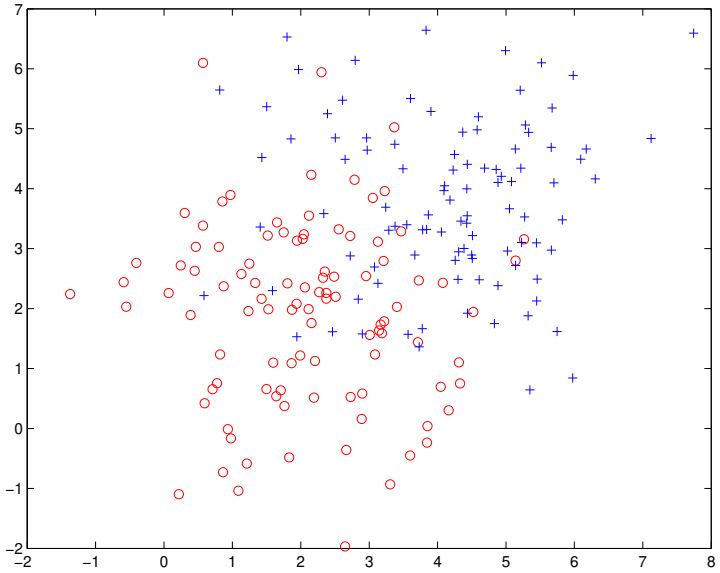


Figura 3.10: Distribuição das Classes

As soluções não-dominadas para este problema estão representadas na Figura 3.11. Na Figura 3.11 está traçada a curva de Pareto em sua escala original (linear) e a curva

de Pareto em escala logarítmica, buscando a forma de um L e com a solução ótima no corner desta curva em L.

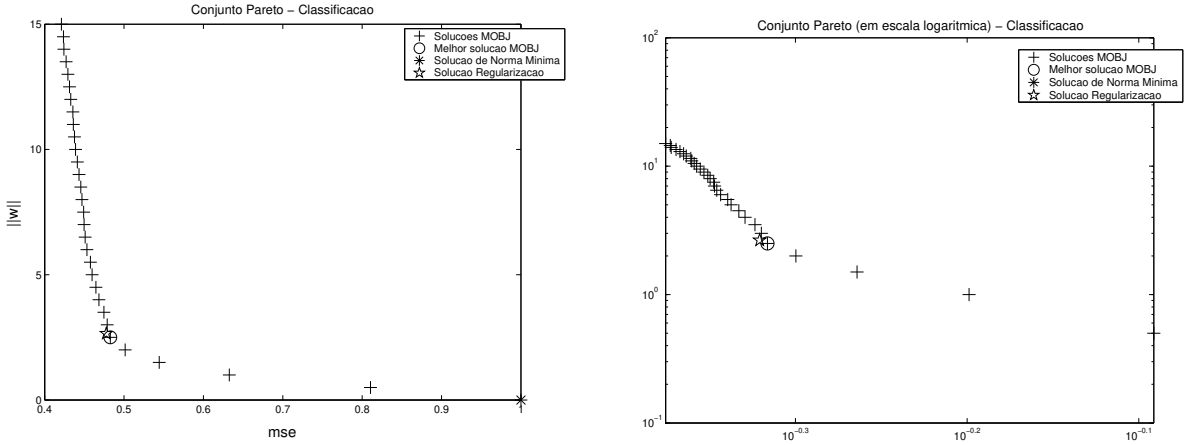


Figura 3.11: O Conjunto Pareto e a *Curva L* (Logarítmica) - Problema Gaussianas.

Na Figura 3.12, está a superfície de decisão obtida pelo método multiobjetivo em RNAs e que está localizada na região do corner da *Curva L* neste problema, conforme mostra a legenda da Figura 3.11. Aqui a curva Pareto, em escala logarítmica, resume-se à *Curva L* e, portanto, o corner da curva em formato L é a própria região em que se encontra a melhor solução multiobjetivo. A metodologia multiobjetivo, sem utilizar a informação geométrica da *Curva L*, encontra a solução situada nessa região de maior curvatura, ou seja, o corner.

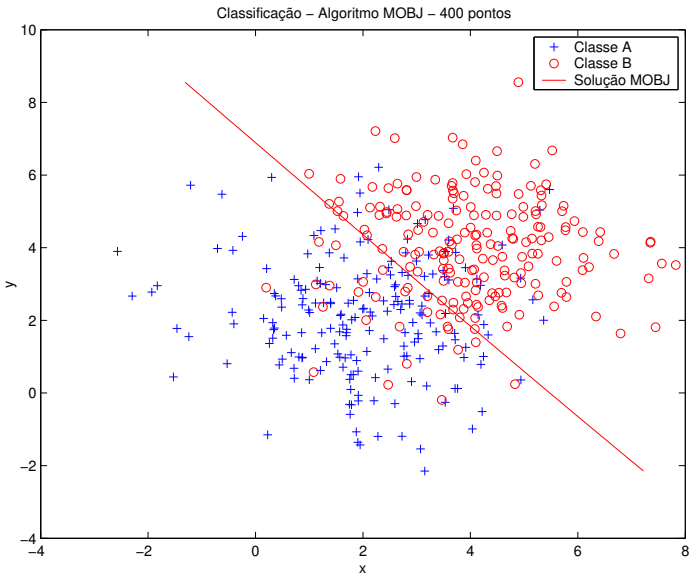


Figura 3.12: Curva de Decisão da melhor solução multiobjetivo, localizada na *Curva L*

Nestes casos, a curva de Pareto e a melhor solução MOBJ (ambas traçadas no espaço dos objetivos em escala logarítmica) resumem-se à *Curva-L*. Portanto, no corner da *Curva L* também está a melhor solução MOBJ, que equilibra os objetivos erro e norma. É possível perceber que, nestes casos, a melhor solução é a de menor distância da solução utópica em um problema multiobjetivo.

Porém, as semelhanças entre a Curva de Pareto e a *Curva L*, resumem-se à abordagem gráfica de exibição do compromisso entre o ajuste e a complexidade das soluções geradas. O método da *Curva L* foi proposto para ser aplicado em técnicas de regularização, encontrando a solução com o melhor parâmetro de regularização (λ) e, tem apresentado a tendência de obter soluções mais regularizadas que o MOBJ para os problemas. Ainda, como foi mostrado em Medeiros (2006), o treinamento de RNAs baseado em técnicas de regularização não é capaz de obter soluções Pareto-ótimas em regiões não-convexas do espaço dos objetivos. Isto é justificado devido à formulação ponderada utilizada em técnicas de regularização, na qual os objetivos são combinados em uma única função objetivo com preferências para cada objetivo em particular. Esta formulação, conhecida como abordagem por somatório dos pesos, discutida em Marler and Arora (2010), não é capaz de mapear pontos em regiões não-convexas da fronteira Pareto-ótima.

Classificação de Padrões - Problema da Espiral

A Figura 3.15 mostra os resultados de um aprendizado de uma RNA através de um conjunto de dados de duas classes gerado por uma função em espiral. A Figura 3.13 mostra a base de dados completa que foi utilizada para o aprendizado.

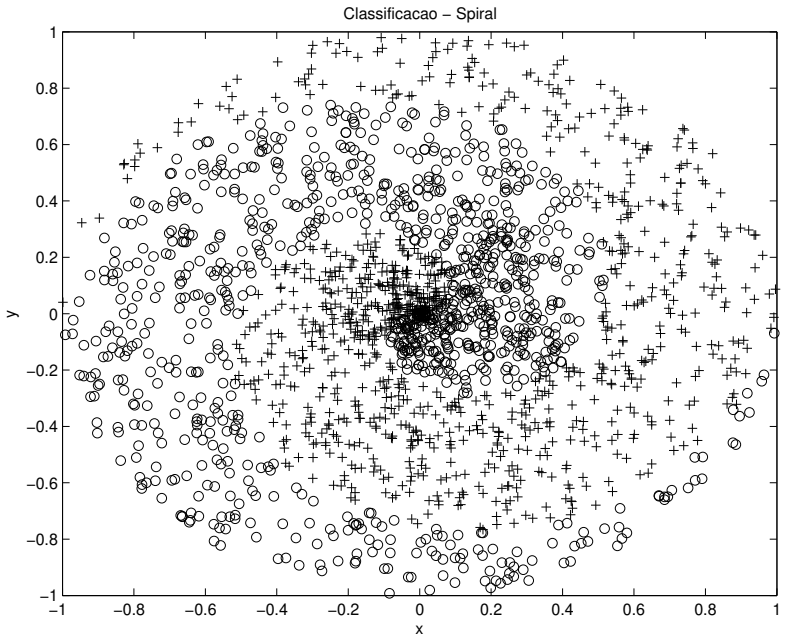


Figura 3.13: Problema de Classificação em espiral

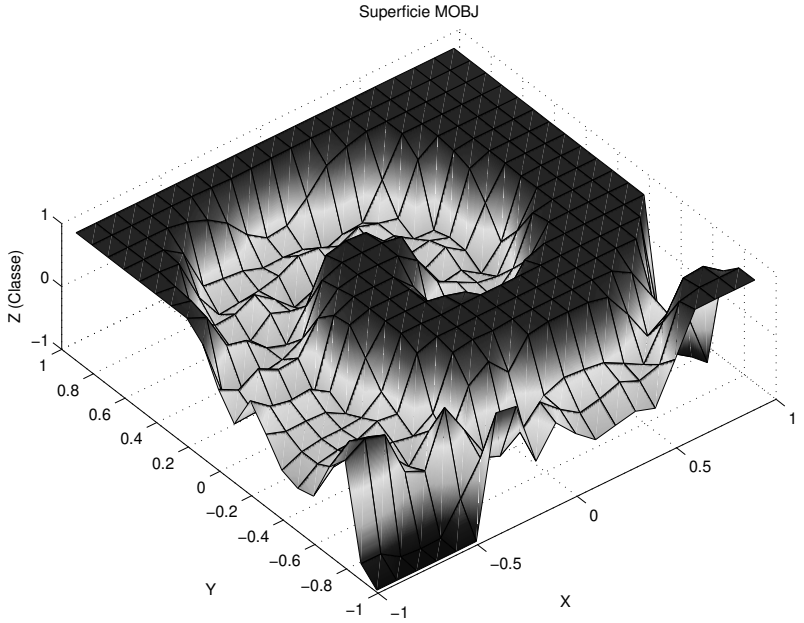


Figura 3.14: Superfície gerada pelo método multiobjetivo

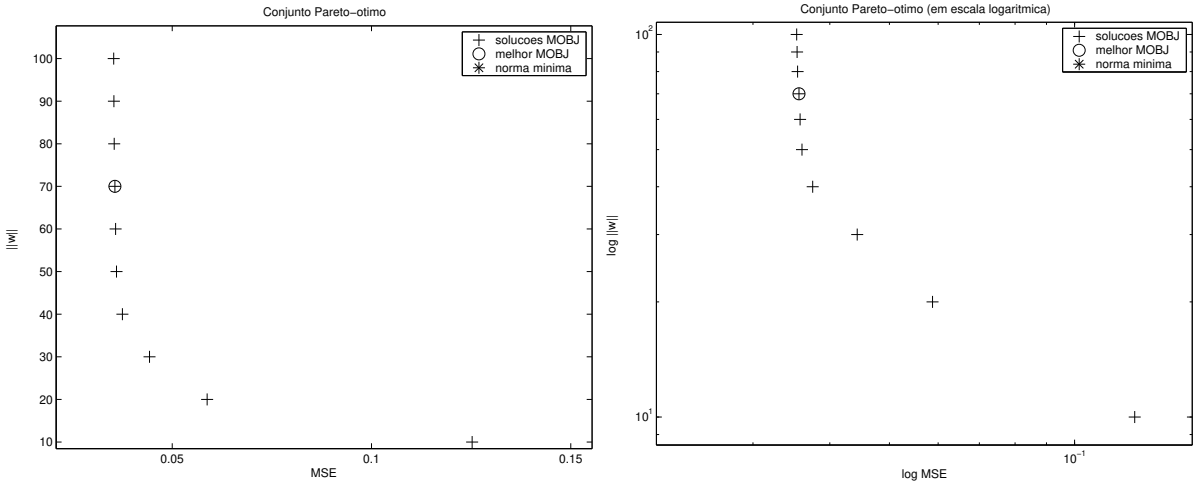


Figura 3.15: Conjunto Pareto e a *Curva L* (Logarítmica) - Problema Espiral.

Nesta simulação, o resultado do método multiobjetivo localiza-se acima do corner obtido com a curva Pareto modificada para a escala logaritmica. Em um problema de aprendizado, onde o critério da *Curva L* seria utilizado, a solução seria regularizada (suavizada) acima do necessário. Isto tornaria a solução sub-parametrizada, ou seja, apresentando uma tendência à polarização da resposta de rede.

Problema de Regressão Não-linear Unidimensional

A Figura 3.17 mostra os resultados de um aprendizado de uma RNA MLP através de um problema de regressão não-linear uni-dimensional. A Figura 3.16 mostra a base de dados completa que foi utilizada para o aprendizado.

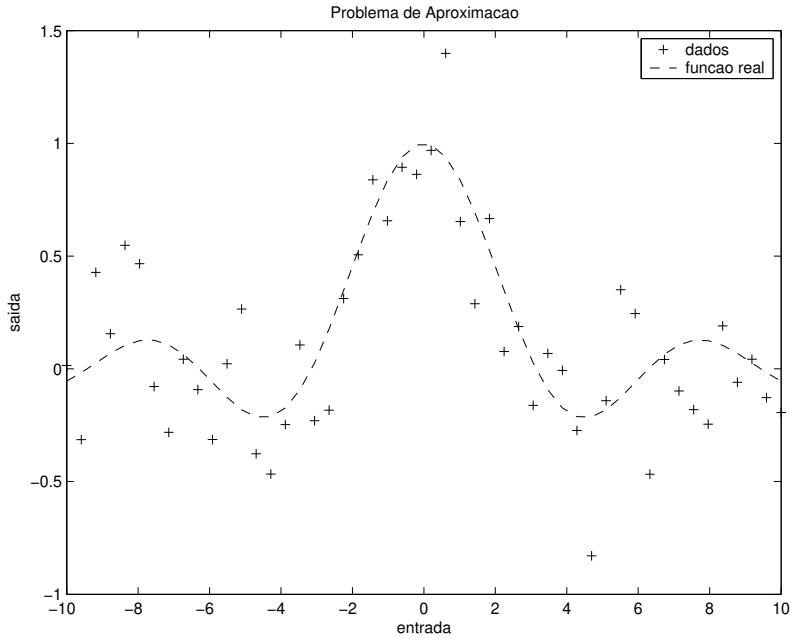


Figura 3.16: Problema de Regressão Não-linear

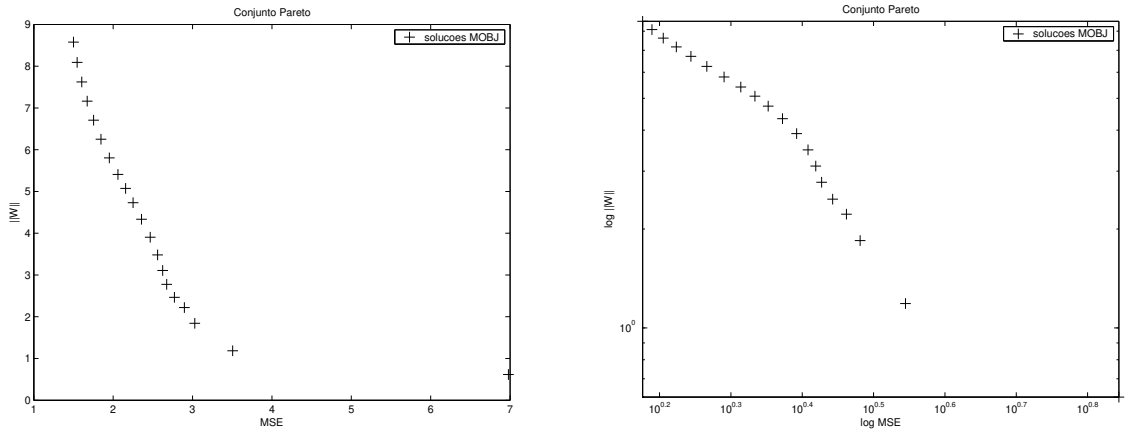


Figura 3.17: Curva de Pareto (Linear) e a *Curva L* (Logarítmica) - Problema Regressão Não-Linear.

A Figura 3.17 mostra que para este problema de aproximação, a curva de Pareto

obtida, ao ser representada em escala logarítmica, não corresponde a uma curva em L, como desejado. Neste caso, a abordagem baseada na *Curva L* para definição da melhor solução não é capaz exibir um corner bem definido, dificultando o processo de decisão. Isto mostra que este critério não consegue ser aplicado indistintamente em problemas de aproximação com a mesma eficiência que o método MOBJ que consegue obter um critério de decisão independente da forma da Curva de Pareto.

3.2 Comentários Finais

Este capítulo apresentou resultados comparativos da análise entre o método multiobjetivo de aprendizado e o método de regularização com a *Curva L*. Os resultados obtidos mostram que ambos os métodos são formalmente semelhantes na determinação de soluções, pois, ambos tratam o erro de ajuste aos dados e a complexidade do modelo como termos de comportamentos opostos. Experimentalmente foi possível verificar que a aplicação do conceito da *Curva L* em redes MLP não garantia o formato característico de "L" e nem com o corner localizado na mesma região da solução escolhida. A diferença está na generalização que o método multiobjetivo possui em problemas de aprendizado independente do modelo, enquanto a formulação da *Curva L* foi proposta inicialmente para o domínio dos modelos lineares, tal como RBFs.

Para os problemas de aprendizado nos quais as RBFs aplicam-se bem, o conceito da *Curva-L* pode ser usado diretamente da construção do modelo que melhor equilibra a polarização e a variância. Portanto, o critério de busca do corner pode também ser usado como decisor no método MOBJ para treinamento de RBFs, Polinômios e outros modelos lineares de aproximação de funções. Existem trabalhos que apresentam o conceito de uma *Curva L* não-linear (Gulliksson and Wedin, 1998; Eriksson and Wedin, 1997), porém apesar dessa nova abordagem, o problema de aproximação de uma função desconhecida continua sendo formulado por uma função de custo baseada em uma combinação linear dos objetivos. Aqui ainda cabe uma investigação mais detalhada sobre as propriedades dessa forma não-linear da *Curva L*.

No entanto, o método de decisão por erro de validação do MOBJ apresenta-se como uma proposta do aprendizado multiobjetivo em que a solução escolhida deve ser consistente para a maioria das situações em que a *Curva L* mostra-se limitada. A decisão por erro de validação, porém, necessita de uma amostra do conjunto de dados do problema para realizar tal decisão. Nesse ponto, a decisão ainda é pouco eficiente. Ainda sobre o decisor por validação, sua escolha costuma ter um viés de escolha por soluções levemente sobre-ajustadas aos dados. Ainda cabe uma investigação mais detalhada para abordar as propriedades do critério de erro de validação buscar efetivamente a variância mínima do resíduo, nos caso de problemas de regressão. Essa observação sobre o decisor por erro de validação não o desqualifica como uma estratégia, apenas faz uma ressalva sobre as soluções por ele apontadas em alguns casos.

Os capítulos seguintes serão destinadas à apresentação de outras estratégias de decisão no conjunto de soluções Pareto-ótimas visando estender a abordagem multiobjetivo de treinamento de máquinas de aprendizagem e relatando o cenário de aplicação dessas

estratégias, bem como suas propriedades.

Capítulo 4

Estratégia de Decisão em Classificação

Neste capítulo é apresentada uma estratégia de seleção de modelos que motivada no procedimento de teste de hipóteses para indicar a solução final para os problemas classificação de padrões. A solução proposta é fundamentada na aproximação da distribuição binomial do processo de classificação de um exemplo. Essa abordagem nos garante a independência da separação das amostras e sua importância em resolver um importante dilema no processo de decisão por modelos de máquinas de aprendizagem Pareto-ótimos. A solução obtida, derivada de uma reinterpretação do princípio da simplicidade é justificada e exemplificada com simulações computacionais em problemas diversos de classificação de padrões.

4.1 Introdução

Os algoritmos de aprendizagem para tarefas de classificação usam em sua maioria a informação de um conjunto de dados de validação para escolher o modelo que melhor represente a função $f_g(.)$ geradora dos dados. As Máquinas de Vetores Suporte são capazes de obter uma aproximação da função geradora por meio do conceito de margem de separação. No processo de decisão multiobjetivo, usa-se frequentemente uma estratégia baseada em uma amostra $\mathcal{T}_V = \{x_i, (d_i+\xi_i)\}_{i=1}^{N_V}$ com novos dados extraídos de acordo com a mesma distribuição D que gerou os dados de treinamento. A máquina de classificação escolhida, com base na hipótese por ela aprendida, deverá apresentar o menor erro para a amostra $\mathcal{T}_V$, denominada por conjunto de validação. Esta estratégia adotada para decisão apresenta resultados satisfatórios, especialmente quando o conjunto de validação é estatisticamente representativo. Porém, esta é uma consideração que nem sempre pode ser garantida e pode levar a decisões por soluções inadequadas.

Em algumas situações, o conjunto de validação $\mathcal{T}_V$ poderá durante o processo aleatório de separação dos dados, apresentar uma tendência no espaço de distribuição, privilegiando certas regiões e, consequentemente, negligenciando outras. A má distribuição

espacial dos dados pertencentes a este conjunto não apresentará as características relevantes da classe.

Além disso, o conjunto de validação pode ser estatisticamente insuficiente para representar o espaço amostral e, dessa forma, não ter informação suficiente que permita ao decisor tomar uma decisão satisfatória. Isto ocorre frequentemente em problemas de grandes dimensões que precisam de um número maior de amostras para manter a mesma densidade da amostra de problemas de dimensões reduzidas e, em princípio, ser suficiente para solução do problema.

Por último, é comum ter-se uma amostra reduzida, devido à dificuldade inerente da coleta de dados de um problema, o aproveitamento de todos os exemplos disponíveis para o processo de ajuste dos parâmetros é muito importante. Uma situação problemática surge quando é necessário separar o conjunto de validação de uma amostra reduzida. Uma outra situação problemática derivada do processo de decisão com base em um conjunto de validação é que o decisor, independente da qualidade dos dados de validação, sempre procura escolher a solução que apresente o menor erro de validação. Apesar de gerar soluções satisfatórias na maioria dos casos, dependendo do conjunto de soluções não-dominadas gerado na etapa anterior do método MOBJ a etapa de decisão pode ser induzida a realizar escolhas com algum grau de sobre-ajuste aos dados.

Apresentadas as situações problemáticas do processo de decisão baseado em um conjunto de validação, torna-se importante a construção de uma nova estratégia que ajude a minimizar o problema da dependência de dados. A solução deste problema é bastante relevante no processo de seleção de máquinas de aprendizagem pertencentes ao conjunto Pareto-ótimo de soluções.

4.1.1 O Princípio do decisor

As principais ideias do decisor para problemas de classificação sob a abordagem multiobjetivo são apresentadas com o objetivo de ressaltar os pontos relevantes da estratégia de decisão.

Os métodos multiobjetivo para classificação buscam o equilíbrio entre a habilidade da máquina de aprendizagem dar respostas corretas para o problema de classificação versus a regularidade de tal máquina. Este raciocínio faz, implicitamente, referência ao princípio da simplicidade (*Occam's razor*) quando escolhe uma hipótese de um conjunto de alternativas possíveis, a preferida é a mais simples.

Entretanto, desconhece-se *a priori*, a partir de um conjunto de dados, se tal conjunto necessita de uma estrutura complexa ou se deveria ser modelado por uma estrutura simples. A metodologia MOBJ original utiliza uma segunda amostra com o objetivo de tomar a decisão no conjunto Pareto-ótimo. Esta metodologia fornece uma forma muito razoável para decidir qual hipótese (solução) escolher. O raciocínio subjacente que está por trás desta metodologia pode ser explicado como: o modelo que produz a hipótese com o menor erro de classificação para a segunda amostra de exemplos deverá ser escolhido. Este procedimento, em princípio, evita a sobre-modelagem que resulta na simples minimização do erro de classificação sobre os dados de treinamento, o que nos leva à modelagem do ruído dos exemplos.

Dessa forma, acredita-se que o raciocínio anterior pode ser reinterpretado como: o modelo correto deve produzir uma proporção de classificações errôneas no conjunto de treinamento semelhante ao erro de classificação verificado em outros dados. A segunda amostra, o conjunto de dados para validação, é utilizado como tais dados.

Partindo da nova declaração, a nova suposição explica, de modo significativo, o que é o princípio absoluto por trás do processo de decisão: ele tenta reproduzir no modelo a taxa de erro que está associada à geração dos dados. Esta interpretação é mais significativa que a interpretação relativa que tem-se utilizado comumente, que sugere que o conjunto de validação é somente uma forma de comparar um conjunto de alternativas de máquinas de classificação.

O resultado mais importante dessa nova interpretação, entretanto, é que é possível suportar o desenvolvimento de outras metodologias de decisão, que diretamente reproduzem o erro estatístico. Tais metodologias, portanto, não precisarão de qualquer processo de separação de amostras, sendo especialmente relevantes e bem vindas em problemas de classificação de amostras pequenas.

A Figura 4.1 mostra um conjunto de dados com duas possíveis classificações de uma máquina de aprendizagem, onde um ajuste produzido não apresenta erros de classificação para o conjunto de dados e o segundo produz uma superfície de separação mais regular. A escolha entre as duas alternativas deve ser realizada de acordo com a nova interpretação proposta, com relação ao conhecimento *a priori* sobre a confiabilidade do conjunto de dados. Se o conjunto de dados é absolutamente confiável (i.e., se ele não apresenta nenhum erro de classificação nos dados originais, significando que os dados não apresentam nenhum ruído) então a máquina de aprendizagem com nenhuma classificação errônea deverá ser escolhida.

Entretanto, se é conhecida a presença de algum ruído no conjunto de dados, então a máquina com nenhuma classificação errônea estará provavelmente errada e a segunda máquina é preferida. Se a taxa de erro de classificação que existe no conjunto de dados tem alguma propriedade estatística conhecida *a priori*, ela possibilitará encontrar qual deverá ser a melhor máquina de classificação, entre um conjunto de possibilidades, com base em tal conhecimento.

A ideia principal, portanto, é aplicar o princípio de um teste estatístico que quantifique a probabilidade de uma dada máquina de classificação ser a melhor, comparada com as outras do conjunto Pareto-ótimo, com base em algum conhecimento *a priori* sobre a distribuição do erro. Com essas estatísticas calculadas, temos um decisor que usará essas informações para realizar a seleção do modelo. É necessário destacar que a decisão deve considerar somente o conjunto reduzido de soluções previamente obtidas no método multiobjetivo. É óbvio que outras máquinas de aprendizagem podem apresentar o mesmo erro de classificação, mas com complexidade diferente.

O exemplo mais simples desta ideia é o caso em que existe uma probabilidade fixa do erro de classificação, não importando a posição da amostra no espaço de características. Supondo que existem duas classes, A e B, e supondo que é conhecido que o gerador de dados produz um erro descrito por:

- Um exemplo que pertence à classe A possui uma probabilidade p_1 de ser rotulado

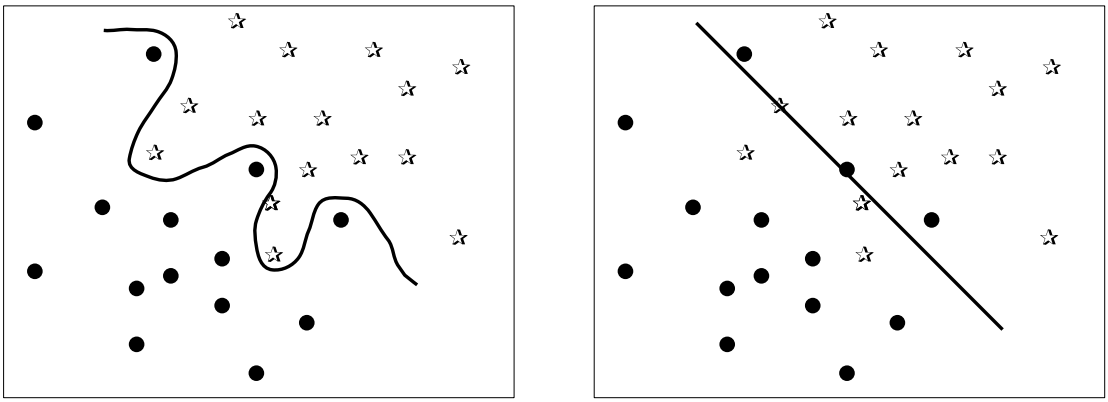


Figura 4.1: Classificações de Máquinas Alternativas: Exata (super-ajustada) e regular (ajustada).

como pertencente à classe B, e um exemplo que pertence a classe B possui uma probabilidade p_2 de ser rotulado como pertencente a classe A.

Neste caso, o número de classificações errôneas em cada classe segue uma *distribuição binomial*.

Uma abordagem ingênua para a tomada de decisão, neste caso, é escolher a máquina de aprendizagem que produzir os desvios empíricos que mais se aproximarem da taxa de erro do conhecimento prévio.

No entanto, uma abordagem mais significativa pode ser formulada da seguinte maneira: dados os desvios empíricos, a probabilidade de aparecerem esses desvios é quantificada, usando o conhecimento da distribuição de probabilidade binomial. A aproximação dessa distribuição indica a probabilidade de cada solução do conjunto Pareto-ótimo, sendo escolhido o modelo mais provável. Em algumas situações, um modelo único iria surgir; em outras situações, alguns modelos podem parecer serem similares; haveria casos em que quase todos os modelos seriam possíveis, configurando um caso sem decisão. Na prática, todos os casos podem de fato ocorrer, configurando as nossas informações padrão. Seria ingênuo afirmar que deve-se ter uma resposta em cada caso individual, não importando quais informações que estejam disponíveis sobre o conjunto de exemplos.

O caso da distribuição binomial nos permite ter um ensaio analítico sobre a tomada de decisão. Se forem introduzidas informações prévias mais sofisticadas, é possível que o princípio do teste de hipótese seja realizado computacionalmente, ao invés de analiticamente.

Ao analisar uma hipótese aprendida (h) sobre um conjunto de treinamento (T) de N amostras extraídas de uma distribuição (D), é possível questionar o quão provável é o erro de treinamento de ser uma estimativa verdadeira do erro de generalização. Ou seja, dado que o conjunto de soluções Pareto-ótimas corresponde a um número reduzido de funções aproximadas, alguma(s) delas apresentará o risco estrutural mínimo, indicando

a solução de melhor capacidade de generalização.

Para explicar melhor os conceitos de erro amostral (erro de treinamento) e o erro provável (erro de generalização) a próxima seção irá abordar as duas noções de precisão úteis para o restante desse capítulo.

4.2 O Erro Amostral e o Erro Provável

O *erro amostral* é a taxa de erro de uma hipótese h sobre um conjunto de dados disponível T de exemplos extraídos de uma população X, ou seja, é a fração de T que h classifica erroneamente.

Definição 4.2.1 (O Erro Amostral) *O erro amostral, dado por $erro_T(h)$ de uma hipótese h em relação à função objetivo f e a amostra de dados T é*

$$erro_T(h) \equiv \frac{1}{N} \sum_{x \in T} L(f(x), h(x)), \tag{4.1}$$

onde N é o número de exemplos de T, e $L(f(x), h(x))$ é uma função de perda indicadora onde a saída é 1 se $f(x) \neq h(x)$ e 0, caso contrário.

O *erro provável* é a taxa de erro de uma hipótese h sobre toda a distribuição desconhecida D de exemplos, ou seja, a probabilidade de que h classificará erroneamente uma única instância extraída aleatoriamente da distribuição D.

Definição 4.2.2 (O Erro Provável) *O erro provável da hipótese h, dado por $erro_D(h)$ em relação à função objetivo f e distribuição D é*

$$erro_D(h) \equiv Pr_{x \in D}[f(x) \neq h(x)], \tag{4.2}$$

onde $Pr_{x \in D}$ denota que probabilidade é considerada sobre a distribuição D.

É esperado obter-se uma estimativa aproximada do erro provável (erro de generalização) $erro_D(h)$ da hipótese, pois este é o erro esperado quando se aplica a hipótese a novos exemplos. No entanto, a informação obtida durante o treinamento é somente o erro amostral $erro_T(h)$ (erro de treinamento) que é um estimador do erro provável . Diante disso, espera-se obter a hipótese que apresente o erro amostral mais representativo do erro provável, dentre as hipóteses aprendidas em cada um dos modelos contidos no conjunto de soluções Pareto-ótimas.

No caso de problemas de classificação onde considera-se um conjunto de dados T com N exemplos extraídos de uma distribuição D, é gerada uma hipótese h definindo uma função indicadora aproximada.

Considerando, ainda, que a hipótese h classifica erroneamente r desses N exemplos ($erro_T = \frac{r}{N}$), então, se não existir nenhuma outra informação,é possível afirmar que o erro amostral $erro_T(h)$ é um estimador do erro provável $erro_D(h)$.

No entanto, é conhecido que para conjuntos de dados reduzidos a estimativa do erro amostral é pouco confiável para representar o erro provável. E, em diversos casos, o conjunto de dados é bastante reduzido.

Porém, é possível construir uma função da distribuição de probabilidades do erro amostral $erro_T(h)$ comportando-se como uma variável aleatória que obedece a distribuição Binomial. Isso é possível graças às características das saídas geradas pelas hipóteses durante o treinamento, correspondentes a uma função indicadora.

Na próxima seção deste capítulo serão discutidos os conceitos básicos da construção de uma distribuição binomial baseada na ocorrência de uma quantidade de erros amostrais em cada hipótese do conjunto Pareto-ótimo.

4.3 A Distribuição Binomial

A distribuição binomial é a distribuição de probabilidade discreta do número de sucessos numa sequência de n tentativas tais que as tentativas são independentes; cada tentativa resulta apenas em duas possibilidades, sucesso ou fracasso (conhecida como tentativa de Bernoulli); a probabilidade de cada tentativa, p, permanece constante (Peebles, 1993).

A função de probabilidade é dada por:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \cdots, n, \tag{4.3}$$

sendo k o número de sucessos e $\binom{n}{k}$ a representação do coeficiente binomial calculado por

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{4.4}$$

A esperança ou o valor médio de X, $E[X]$ é dado por:

$$E[X] = np \tag{4.5}$$

A variância de X, $Var(X)$ é dada por:

$$Var(X) = np(1-p) \tag{4.6}$$

Seguindo o princípio fundamental da construção de uma distribuição binomial, foi proposta uma adaptação do uso da distribuição para representar as probabilidades das hipóteses do conjunto Pareto-ótimo indicarem o erro provável para cada classe do problema de classificação. Assim, em um problema de aprendizagem composto por classes, cada hipótese terá sua probabilidade de representar o erro provável em cada classe.

Dessa forma é bastante plausível que esta abordagem de decisão não indique apenas uma solução, mas uma região constituída de soluções equivalentes. Isto é possível porque diferentes hipóteses poderão ser mais prováveis em classes diferentes, indicando que as hipóteses intermediárias são equivalentes.

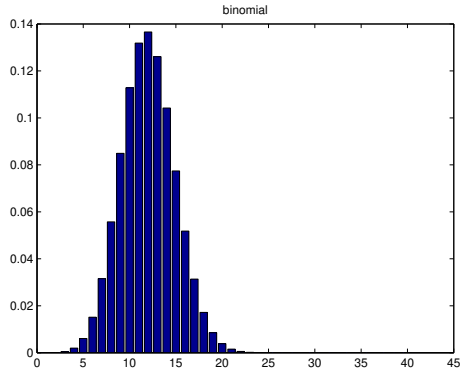


Figura 4.2: A Distribuição Binomial para $n = 40$ e $p = 0.3$.

O ponto chave dessa estratégia de decisão é a forma como o conhecimento prévio do problema é incorporado ao processo de decisão. A informação de classificações errôneas (p) no conjunto de exemplos de treinamento é usada para direcionar melhor a busca pela hipótese mais provável.

Dessa forma, a função da Equação 4.3 pode ser reinterpretada da seguinte maneira. A probabilidade $P_j(erros_S(h_i))$ de uma dada hipótese h_i apresentar $erros_S$ classificações errôneas para a classe j é indicada pela Equação 4.7. A Figura 4.2 indica, no eixo horizontal, a probabilidade da ocorrência de cada evento (hipótese) dado o conhecimento prévio p.

$$P_j(erro_S(h_i)) = \binom{N}{erro_S(h_i)} p^{erro_S(h_i)}(1-p)^{N-erro_S(h_i)}, \tag{4.7}$$

onde $erro_S(h_i) = 0, 1, 2, \cdots, N$, e j é o índice da classe em questão.

Cada solução Pareto-ótima terá uma probabilidade medida para cada uma das classes do problema em questão. A decisão final por uma solução é tomada pelo modelo mais provável de representar todas as classes do problema.

Com base nesse princípio, a tomada de decisão no conjunto de soluções Pareto-ótimas, feita com o auxílio da informação de um especialista, irá garantir que a solução indicada pelo decisor será a mais provável de apresentar o menor erro de generalização, ou seja, o erro provável que a máquina de aprendizagem busca.

4.4 Comentários Finais

Neste capítulo foi descrito o princípio básico da formulação de uma estratégia de decisão multiobjetivo para os problemas de classificação. O fator de destaque para a nova abordagem é sua independência do processo de separação de amostras para treinamento e validação. Isso garante que o processo de treinamento da máquina de aprendizagem terá um conjunto de dados mais representativo. Como alguns problemas de classificação

podem apresentar alguma informação prévia da qualidade do conjunto de dados, permite-se usar tal informação na orientação pela busca da melhor solução. Neste caso, foi usada a probabilidade de classificações errôneas nos rótulos de cada exemplo como informação útil e que, até então, não era usada diretamente no critério de decisão.

Com o uso dessa informação *a priori*, tornou-se possível obter uma função aproximada mais coerente em relação à função real, aproveitando de forma mais eficiente os conjuntos de dados reduzidos.

É preciso esclarecer que não é realizado um teste das hipóteses de cada um dos modelos gerados, mas uma avaliação desses modelos através de uma métrica que foi obtida por meio de um princípio de um teste de hipóteses. Logo, o decisor por si só não é o teste de hipóteses, como conhecido na literatura.

Capítulo 5

Simulações da Decisão em Classificação

Neste capítulo são apresentados os resultados da aplicação da estratégia de decisão probabilística em modelos obtidos pelo método MOBJ em experimentos computacionais de problemas de classificação de padrões. Os experimentos foram realizados com duas bases de dados artificiais e uma base com dados reais sobre o problema de diagnóstico médico sobre doenças do coração (Asuncion and Newman, 2007). Os resultados obtidos por este decisor serão analisados com o decisor por erro de validação, com uma Máquina de Vetores Suporte (*Support Vector Machines - SVM*) e, ainda, com uma análise de sua localização na região de máxima curvatura (ou região de menor ângulo agudo) da *Curva L*.

Os gráficos das curvas de separação irão destacar as soluções Pareto-ótimas escolhidas e a solução da SVM. A *Curva L* será usada para destacar a região indicativa das soluções de melhor capacidade de generalização. As probabilidades das hipóteses geradas no método MOBJ para cada classe são apresentadas em gráficos separados, indicando a hipótese mais provável para o problema.

5.1 Problema de Duas Distribuições Gaussianas

Esta base de dados foi gerada artificialmente a partir de duas funções de distribuição gaussianas bi-dimensionais com os respectivos centros nas coordenadas $(1, 1)$ e $(-1, -1)$ e com um nível de ruído que gerasse uma sobreposição dos dados de cada classe.

A base possui 200 exemplos divididos igualmente em duas classes e foi usada para analisar o aprendizado de uma rede neural artificial por meio do método MOBJ com decisão probabilística e por erro de validação. Além disso, os resultados obtidos serão confrontados com uma SVM treinada com a mesma base de dados.

Os experimentos foram realizados com implementações em Matlab®, usando algumas rotinas do *Neural Networks Toolbox*. A rede neural usada é a MLP com 2 neurônios na camada de entrada, 10 neurônios na camada oculta e 1 neurônio na camada de saída. O algoritmo de treinamento usado foi o algoritmo de otimização Levenberg-Marquardt com

restrição de norma dos pesos (Costa et al., 2007). O conjunto de soluções Pareto-ótimas gerado é constituído de 21 alternativas, dentre as quais uma será escolhida. As rotinas usadas e adaptadas para implementação do treinamento de uma SVM estão presentes no *toolbox* de Bioinformática do Matlab®, dentro de um conjunto de algoritmos sobre aprendizagem estatística. As análises seguintes referem-se aos experimentos realizados para destacar as características do decisor.

5.1.1 Análise da Probabilidade *a Priori* do Decisor

O decisor baseado em um teste de hipótese analisa o conhecimento prévio p, uma probabilidade de classificações errôneas existentes no conjunto de dados. Para isso, nesta sessão, será analisada a influência da variação do parâmetro p na decisão.

Os experimentos realizados tiveram o parâmetro p definido com os seguintes valores: 0.03, 0.05, 0.1, 0.2 e 0.5, onde a variação deste parâmetro define preferência por modelos mais super-ajustados aos dados quando p se aproxima de $p = 0.03$ e a preferência por modelos mais sub-ajustados quando se aproxima de $p = 0.5$.

Os gráficos das Figuras 5.1, 5.3, 5.5, 5.7 e 5.9 mostram a curva de separação obtida pelos decisores do método MOBJ e de uma SVM; o conjunto Pareto-ótimo de soluções obtidas pelo treinamento com o conjunto de dados completo (disponível quando é usado o decisor probabilístico); e os gráficos das binomiais obtidas a partir do treinamento multiobjetivo com a indicação dos modelos mais prováveis para os dados disponíveis.

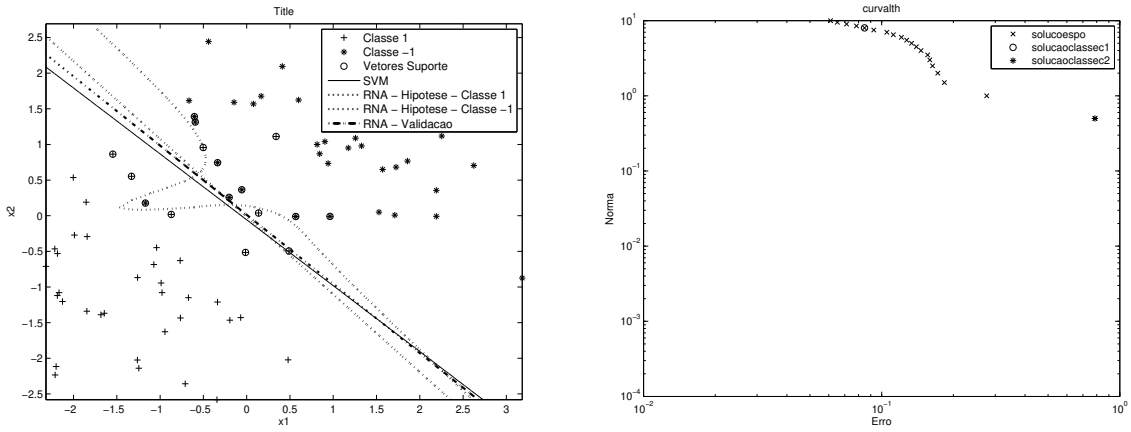


Figura 5.1: Curvas de Separação e *Curva L* gerada com dados do treinamento além da escolha com decisor probabilístico com $p = 0.03$.

Analisando a Figura 5.1 viu-se que cada classe apresentou uma rede mais provável bem distinta da outra, definindo uma faixa muito ampla de soluções intermediárias. Ainda pela Figura 5.1, é possível perceber a diferença entre as decisões. Situações como essa mostram que o valor p definido para tomada de decisão não é adequado. O princípio da decisão por teste de hipótese é obter uma solução ou uma margem de soluções, sendo esta margem não muito ampla para não transparecer indecisão. Os experimentos realizados com valores

ainda menores que $p = 0.03$ mantiveram o mesmo padrão de comportamento. Finalmente, a região de máxima curvatura indicada pela *Curva L* define, para esse critério, a região da solução escolhida.

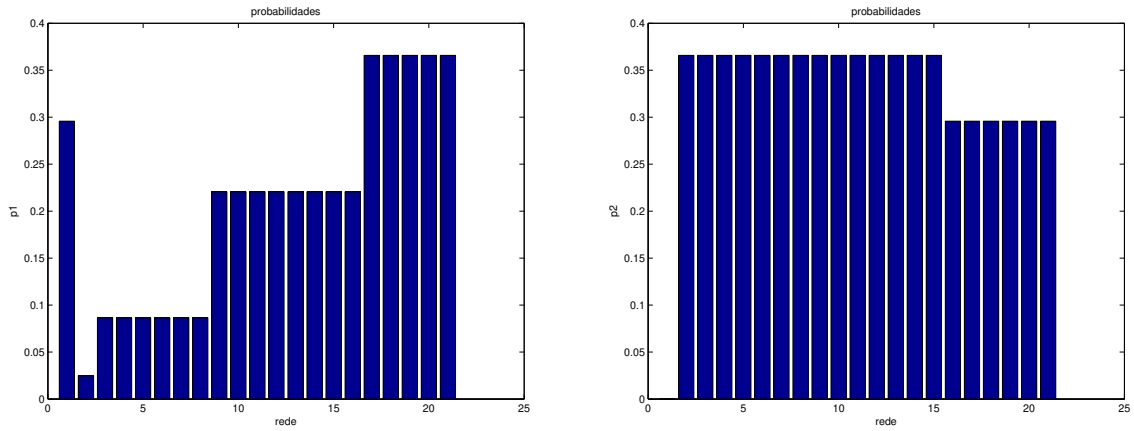


Figura 5.2: Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.03$.

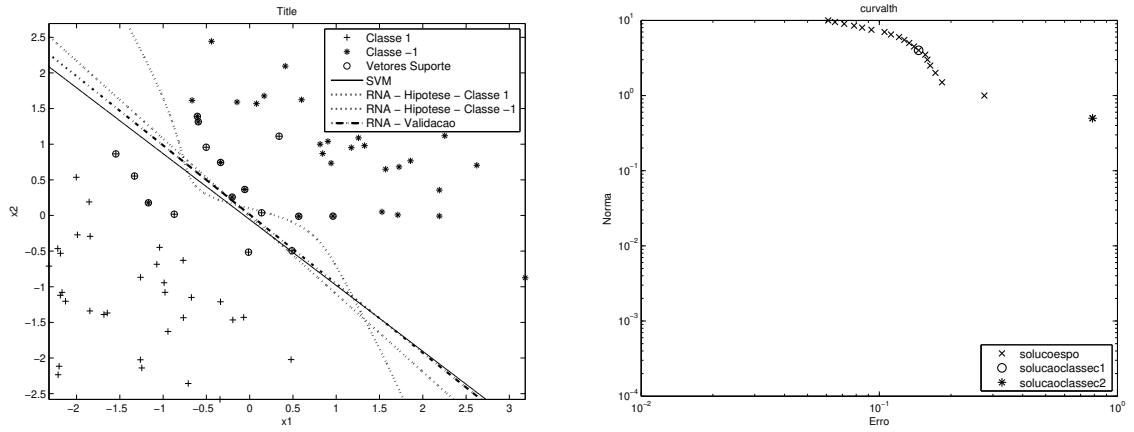


Figura 5.3: Curvas de Separação e *Curva L* gerada com dados do treinamento além da escolha com decisor probabilístico com $p = 0.05$.

Após o aumento do parâmetro p do decisor probabilístico para $p = 0.05$, as soluções indicadas na *Curva L* aproximaram-se um pouco mais. O reflexo dessa aproximação fica mais evidente na curva de separação da Figura 5.3. A Figura 5.4 mostra a região de pico das distribuições geradas, na qual os picos de cada classe estão mais próximos que os apresentados na Figura 5.2. Essa distância tende a diminuir à medida que o valor p começa a privilegiar soluções com maior controle de complexidade.

As curvas de separação obtidas pelo decisor probabilístico com valor $p = 0.10$ apontam para uma única solução do método MOBJ para ambas as classes. Devido a este fato, a

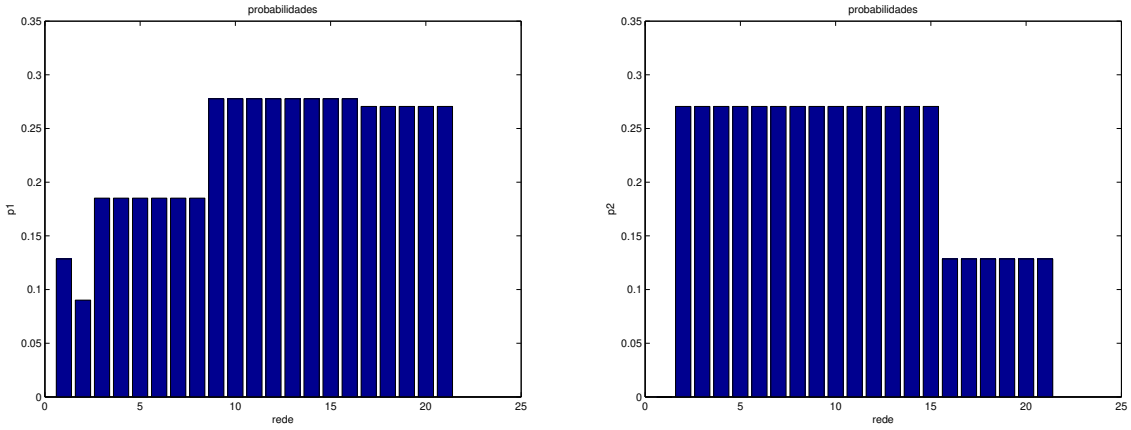


Figura 5.4: Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.05$.

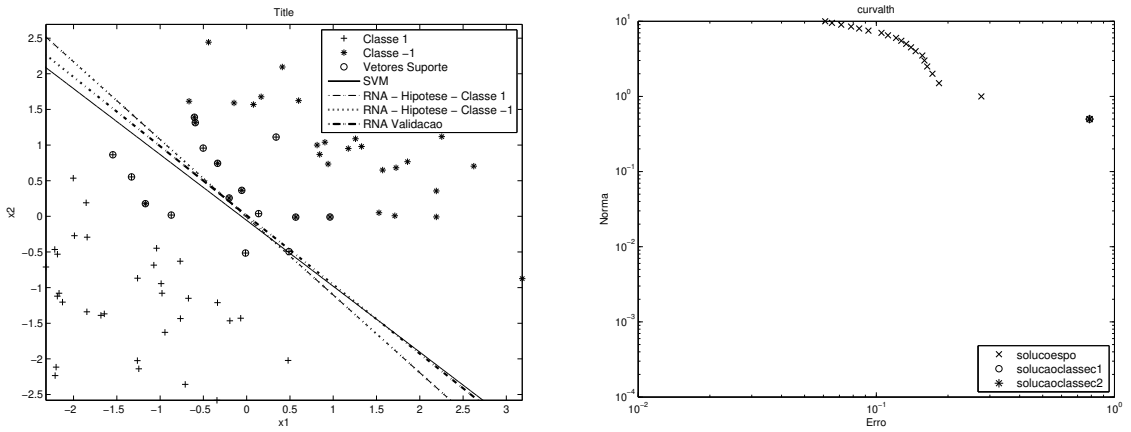


Figura 5.5: Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.1$.

curvas aparecem sobrepostas na Figura 5.5. As curvas do decisor por erro de validação e da SVM, ficam muito próximas uma da outra e, consequentemente, da curva do novo decisor. A Figura 5.6 destaca a hipótese mais provável com as distribuições geradas para cada classe.

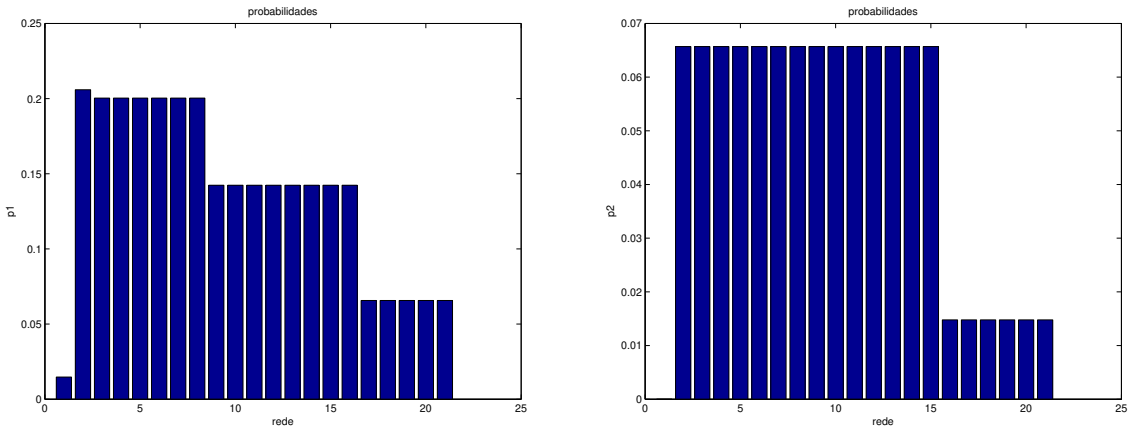


Figura 5.6: Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.1$.

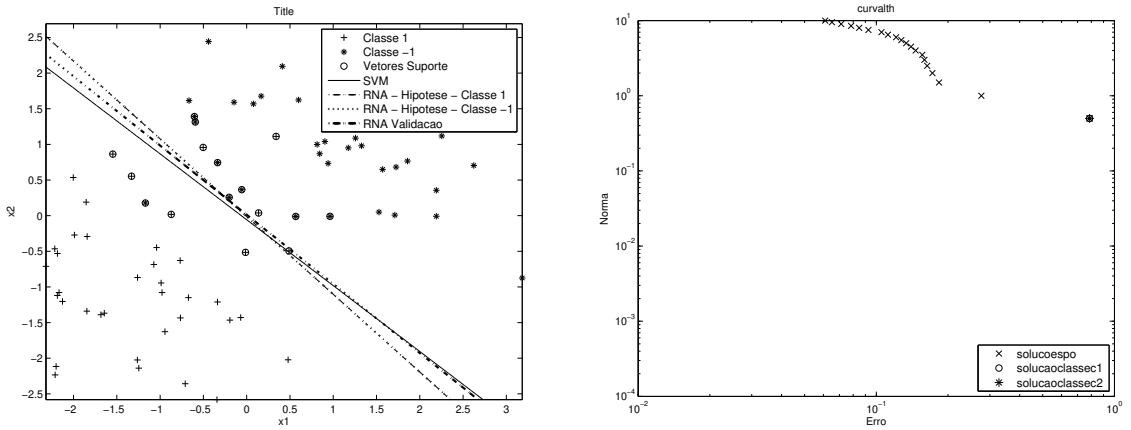


Figura 5.7: Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.2$.

Nas Figuras 5.7 e 5.9 as soluções Pareto-ótimas do decisor probabilístico não sofreram mudanças com o aumento do valor p. Isto garantiu que mesmo com um conhecimento pouco preciso sobre o nível de confiança nos rótulos dos exemplos, o decisor ainda mantivesse uma solução coerente para o problema.

A variação da probabilidade *a priori* usada para obter a distribuição binomial das redes mais prováveis mostrou que pequenos valores indicam uma tendência por obter os modelos mais sobre-ajustados aos dados de treinamento. Em contrapartida, a escolha

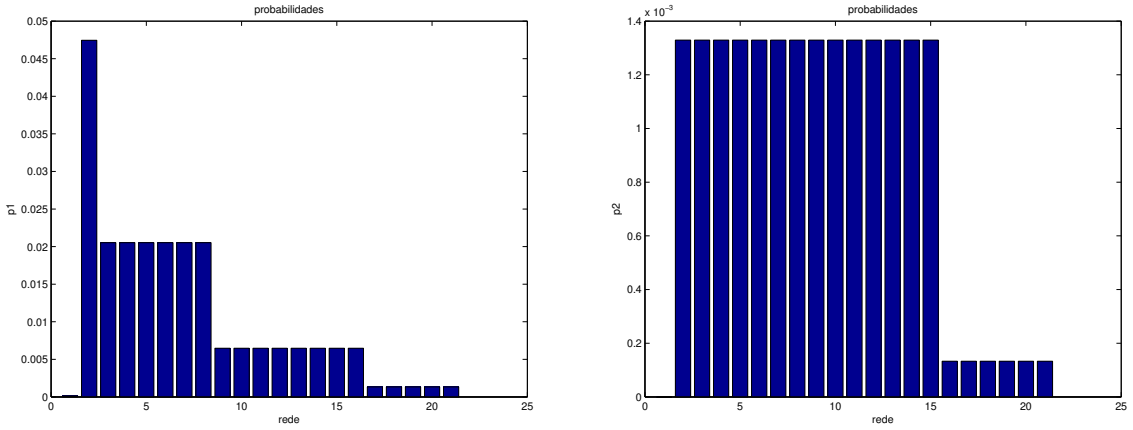


Figura 5.8: Distribuições Binomiais com a indicação da rede mais provável de cada classe usando $p = 0.2$.

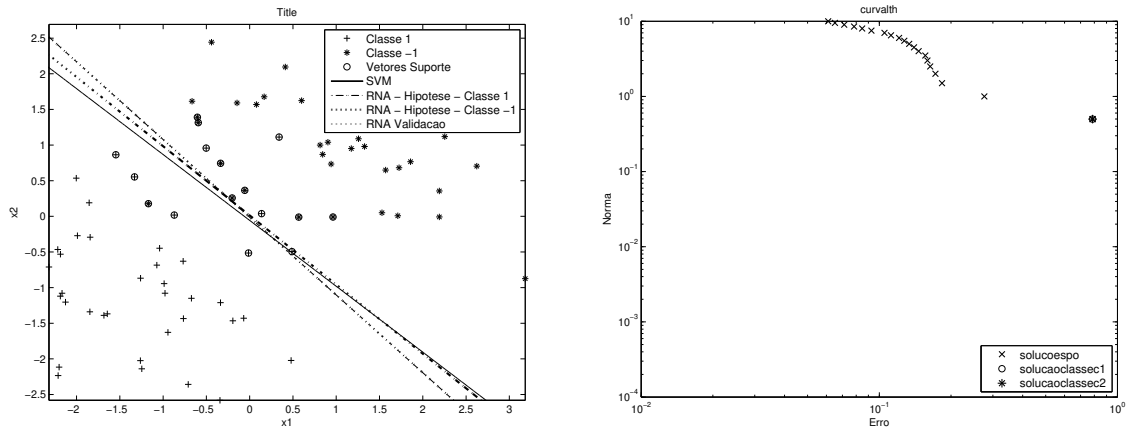


Figura 5.9: Curvas de Separação e *Curva L* gerada com dados do treinamento e escolha do decisor probabilístico com $p = 0.5$.

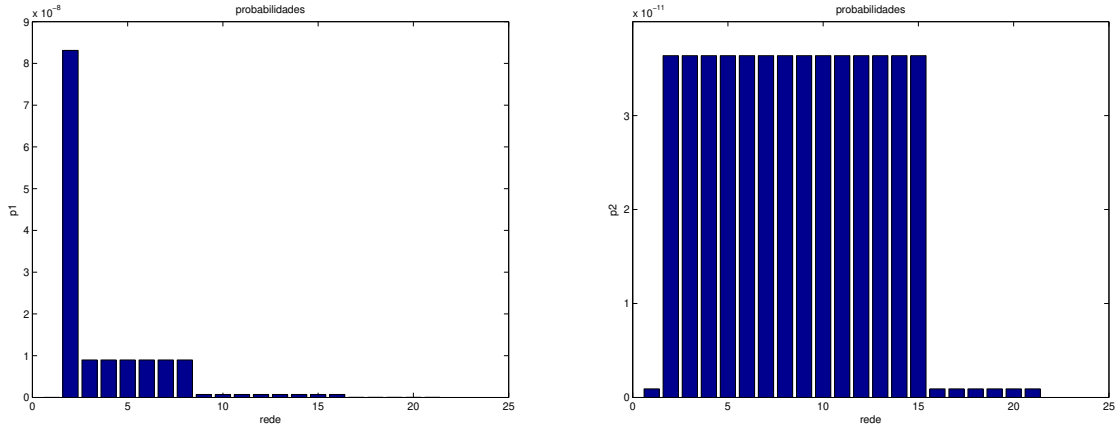


Figura 5.10: Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.5$.

de uma probabilidade *a priori* elevada indicou a escolha os modelos mais sub-ajustados, com maior controle de complexidade. Porém, os experimentos mostraram, para esse problema, que a decisão manteve-se robusta mesmo diante de variações significativas deste parâmetro. Além disso, valores muito próximos de zero para o parâmetro p indicavam que redes muito distintas eram escolhidas para representar cada uma das classes do problema. Assim sendo, esses valores pouco ajudaram no processo de tomada de decisão, indicando uma faixa muito ampla de soluções possíveis para representar ambas as classes do problema de aprendizado.

A SVM apresentou resultados mais precisos somente quando é escolhida a função de kernel adequada para o problema, neste caso, a linear. A escolha de um kernel polinomial não conseguiu obter o controle de complexidade que resumisse o kernel polinomial em um grau tal que ficasse quase linear. A RNA treinada com o conjunto de treinamento e selecionada por um outro conjunto de validação sofreu interferência da densidade do conjunto de dados usado para o treinamento.

A tendência indicada por estes experimentos mostrou que o conhecimento *a priori* pode indicar uma inclinação para soluções com alta polarização (valores p elevados) ou uma inclinação para soluções com alta variância (valores p pequenos). O conhecimento prévio do valor p, que indicará a solução que melhor represente o conjunto de dados, é a principal contribuição desta estratégia de decisão. Pois, uma vez que essa informação é conhecida e permite-se inseri-la no decisor, obtém-se uma estratégia para tomada de decisão que efetivamente busca representar a função geradora dos dados, descartando o ruído inerente.

A variação de p destaca a força dessa informação em direcionar o processo de decisão por modelos com maior ou menor ajuste aos dados disponíveis. E ainda mostrar, por meio da distribuição binomial, a probabilidade de cada modelo representar os dados com o nível de confiança definido. Essa abordagem permite definir uma margem de confiança onde podem ser caracterizadas soluções aceitáveis.

A Escolha

Uma vez que a proposta deste decisor sugere soluções distintas para cada classe do conjunto de dados, a discussão sobre a resposta final do aprendizado multiobjetivo com o decisor probabilístico passará pela análise do quão plausíveis são as regiões das soluções no espaço dos objetivos. Então a sugestão inicial nesta tese é que fique indicada as duas soluções distintas, mas que as regiões por elas definidas sejam analisadas. Torna-se possível a configuração de uma região (i) em que somente uma das classes prevalece segundo a indicação do decisor; uma região (ii) em que ambas as classes são representadas pela(s) solução(ões) indicadas pelo decisor. Esta pode ser considerada uma região de incerteza. Por fim uma região (iii) em que nenhuma classe pode ser representada, sendo esta uma região não coberta por nenhuma escolha do decisor. Essa forma de analisar as escolhas do decisor permite uma discussão mais aprofundada sobre o quão específico ou sensível são as respostas geradas pelo método MOBJ.

Ainda há também uma forma mais simples de se indicar uma solução única para efeitos de implementação rápida de uma solução. Uma vez que o o decisor pode apontar uma das diversas soluções que maximizem simultaneamente as probabilidades de ambas as classes.

Comparação com a curva esperada

Agora, o objetivo dos experimentos com o decisor probabilístico é verificar a estabilidade dessa estratégia para dois cenários bem distintos de amostragem. O treinamento das RNAs pelo método MOBJ foi realizado com conjuntos de dados de 10 e 1000 exemplos. Desse modo, é possível realizar uma análise do comportamento dos decisores no treinamento multiobjetivo em casos extremos, ou seja, com poucos exemplos e com muito exemplos. Além disso, a solução elaborada pela SVM é apresentada e comparada com as soluções do método MOBJ.

Ao realizar o treinamento multiobjetivo com 10 e 1000 dados, obtém-se a curva de separação do modelo escolhido em cada estratégia de decisão. As Figuras 5.11 e 5.12 representam as curvas de separação dos decisores probabilísticos e por erro de validação comparadas com a curva de decisão obtida por uma SVM de kernel linear, treinada com o mesmo conjunto de dados. Apenas no treinamento multiobjetivo com validação foi necessário o fracionamento dos dados em conjunto de treinamento (70% dos exemplos) e conjunto de validação (30% dos exemplos).

Para uso do decisor probabilístico, foi usada a informação de que no conjunto de dados, existem 10% de classificações errôneas em ambas as classes. Essa informação reflete a realidade uma vez que os dados foram gerados artificialmente e, ainda, foram perturbados por um ruído que modificasse cerca de 10% dos rótulos das classes. A distribuição binomial obtida com essa informação definiu as probabilidades de cada uma das redes Pareto-ótimas serem as mais prováveis para cada classe.

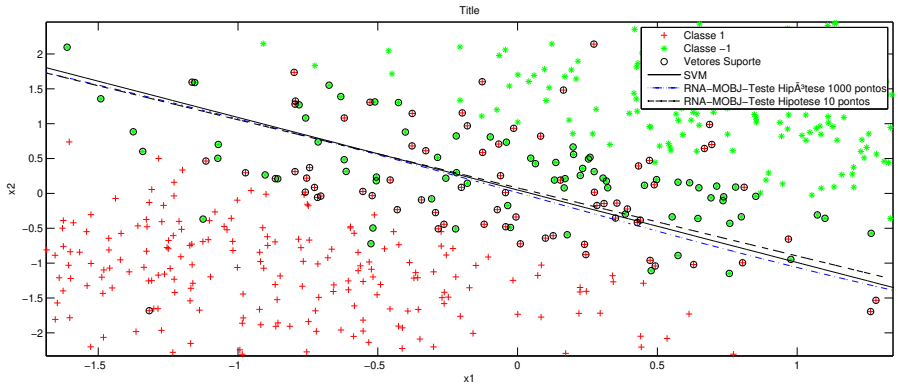


Figura 5.11: Curvas de separação: Decisão probabilística com 10 e 1000 exemplos X SVM treinada com 1000 exemplos.

Na Figura 5.11 as curvas de separação obtidas pela decisão probabilística para 10 exemplos e 1000 exemplos, são bastante semelhantes entre si. A curva de separação da SVM, apresentada na forma de uma linha contínua, fica próxima às outras duas curvas de separação. As curvas de separação obtidas são bem semelhantes entre si, mesmo sendo geradas a partir de circunstâncias bem distintas. Há quase uma sobreposição as curvas de separação geradas pelos métodos. Este era um efeito esperado, uma vez que este

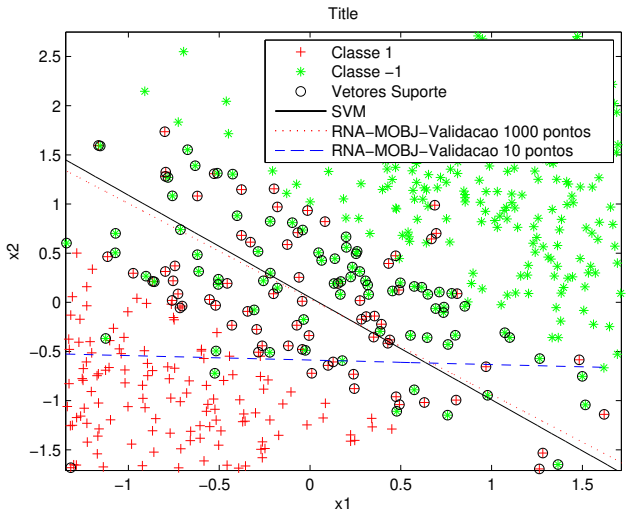


Figura 5.12: Curvas de separação: Decisão por Validação com 10 e 1000 exemplos X SVM treinadacom 1000 exemplos.

processo de decisão é independente de uma amostragem e, portanto, baseado somente em um conhecimento prévio. Neste caso, as bases de dados fornecidas para treinamento das MLPs eram bem diferentes em quantidades.

Na Figura 5.12 a curva de separação obtida na decisão por mínimo erro de validação para 10 exemplos (linha segmentada) fica bastante distante das curvas obtidas pela decisão por mínimo erro de validação com 1000 exemplos (linha pontilhada) e da SVM (linha contínua). A curva de separação fornecida pela MLP, segundo o critério de erro de validação, permite uma instabilidade do decisor, uma vez que o mesmo oscila conforme a qualidade dos dados de validação. Como o conjunto de dados é reduzido, sua representatividade é menor que em um grande volume de dados.

Uma análise do comportamento verificado neste experimento mostra a tendência da alta variabilidade da estratégia de decisão por mínimo erro de validação para conjuntos de dados reduzidos e, também, para amostras com distribuição tendenciosa.

A seguir, a Figura 5.13 exibe uma aproximação da distribuição binomial representando as chances de cerca de 10% de classificações errôneas ocorrerem nos exemplos das classes (+) e (*), respectivamente, com 1000 exemplos no conjunto de dados de treinamento.

A Figura 5.13 mostra que as probabilidades das 9 primeiras redes, não diferem muito e são as mais plausíveis para o conjunto de dados da classe (+). Ainda na Figura 5.13 mostra-se que as 3 primeiras redes são as mais plausíveis, juntamente com a rede 6 e 10 para a classe (*). Porém, as demais redes são bem menos prováveis de representar a classe (*). Dentre as redes com probabilidades iguais, escolhemos, usando o princípio da simplicidade, a rede de menor complexidade.

Em situações nas quais as redes mais prováveis para cada classe apresentem respostas

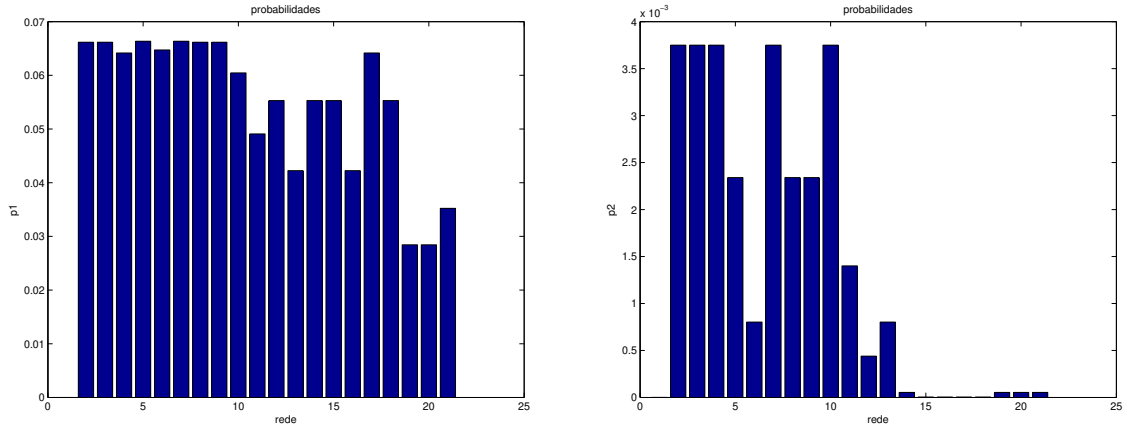


Figura 5.13: Redes mais prováveis de representar as classes +1 (+) e -1 (*), respectivamente.

bem distintas, a rede única não fica caracterizada para o problema. Assim, o intervalo que consiste em tais redes deve ser visto como uma faixa de equivalência entre essas redes. Em alguns casos, quando a solução do Pareto era diferente, mas próxima uma da outra, obtém-se uma solução média que caracteriza a solução escolhida por decisão probabilística. Por exemplo, a rede mais provável para modelar a classe (+) possui $||w|| = 3.0$ e a rede mais provável para modelar a classe (*) possui $||w|| = 5.0$, portanto, escolheu-se a solução média, a rede com valor $||w|| = 4.0$ como solução que equilibra ambas tendências para cada classe.

O método usado para decidir pela rede representativa da decisão para todas as classes não foi considerado único e muito menos o mais eficiente para esta abordagem de estratégia. A princípio, a estratégia de decisão probabilística foi usada para mostrar o quão plausível cada modelo foi para o problema. Quando as distribuições binomiais indicarem soluções muito diferentes entre si, é possível que cada modelo necessite de probabilidades *a priori* diferentes para avaliar separadamente cada classe do problema.

A seguir, a Figura 5.14 exibe o conjunto Pareto-ótimo (em escala logarítmica) com a decisão probabilística para as amostragens de 10 e 1000 exemplos, respectivamente.

A Figura 5.14 indica, por meio da escala logarítmica, uma região de curvatura em L. Nessa região, segundo o princípio da *Curva L*, estaria a solução com melhor equilíbrio entre os objetivos de mínimo erro e mínima norma. Porém, usando uma amostra de 10 exemplos, não fica evidente o *corner* da *Curva L*. No entanto, a amostra de 1000 exemplos permitiu uma aproximação da *Curva L*, onde a região do *corner* pode ser melhor identificada. Neste último caso, o critério da *Curva L* destaca a região de máxima curvatura com soluções de complexidade inferior à indicada pelo decisor probabilístico.

A Figura 5.15 exibe a aproximação da curva Pareto-ótima, em escala logarítmica, usando o método de amostragem baseado em um conjunto de treinamento e de validação. O comportamento verificado nesse método foi o mesmo quando todo o conjunto de dados estava disponível para o treinamento. Para poucos dados de treinamento a curva

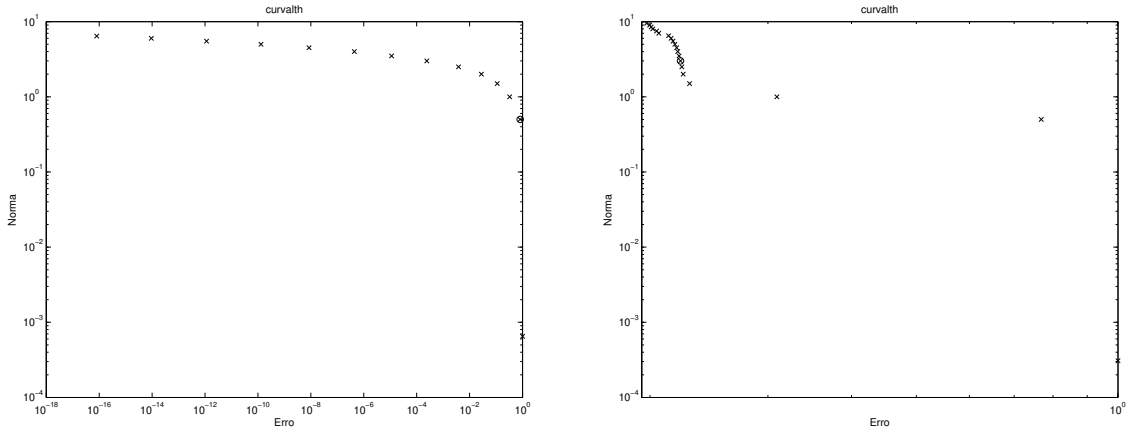


Figura 5.14: Conjunto Pareto, em escala logarítmica, dos modelos obtidos com conjuntos de 10 e 1000 exemplos e decisão probabilística.

não aproximou-se da *Curva L*, enquanto para muitos dados a curva apresentou uma aproximação razoável da *Curva L*.

A estratégia de decisão por mínimo erro de validação escolheu a solução de norma mínima para esta situação. A decisão em um conjunto maior escolheu uma solução super dimensionada na *Curva L*. Neste caso, a solução com norma elevada comporta-se de modo excessivamente variável, apresentando o efeito de *overfitting*.

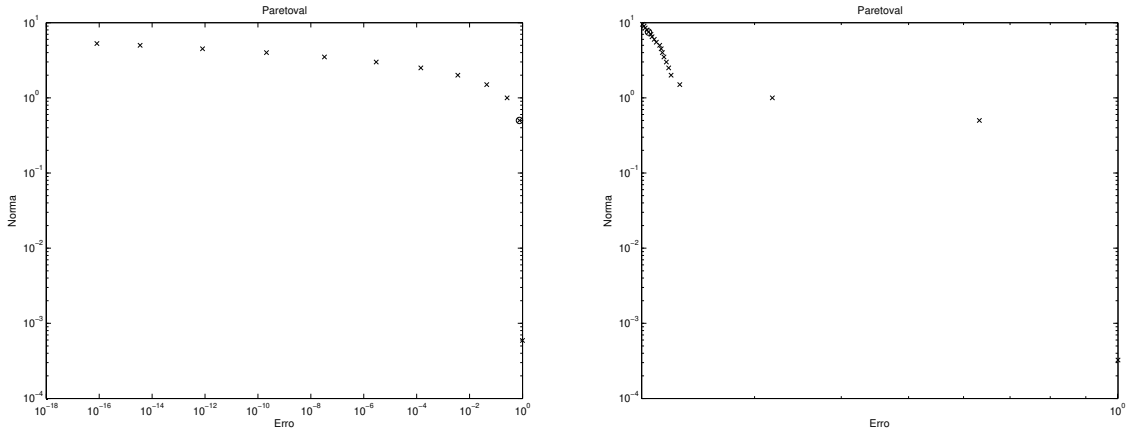


Figura 5.15: Conjunto Pareto, em escala logarítmica, dos modelos obtidos com conjuntos de 10 e 1000 exemplos e decisão por Validação.

5.1.2 Análise do Desbalanceamento das Classes

Nesta seção é analisada a influência de um treinamento com classes desbalanceadas nas estratégias de decisão por erro de validação e probabilística. Além disso, é apresentado o resultado de uma SVM com kernel linear. Aqui 70% do exemplos pertenciam a uma classe -1 e os 30% restantes a classe +1. O decisor probabilístico foi parametrizado com p=0.005, 0.01, 0.02, indicando que 0.5%, 1% e 2% de classificações errôneas deveriam estar presentes nos dados disponíveis.

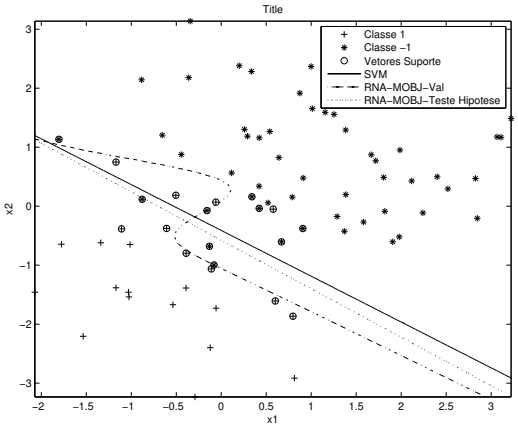


Figura 5.16: Curvas de separação: Decisão por Validação apresenta *overfitting* e p = 0.005.

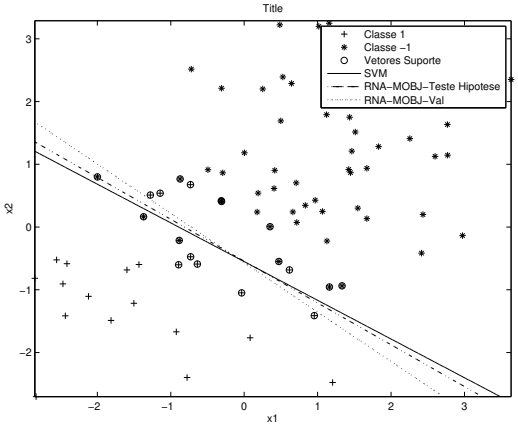


Figura 5.17: Curvas de separação: Decisão por Validação e por Hipótese, com p = 0.01.

Os resultados do treinamento com classes desbalanceadas destacou a eficiência da SVM e do método de decisão probabilística diante do decisor por erro de validação. A Figura 5.18 evidenciou a ineficiência da decisão que está dependente de um conjunto de dados de validação, uma vez que a solução apresentada foi prejudicada pela tendência do

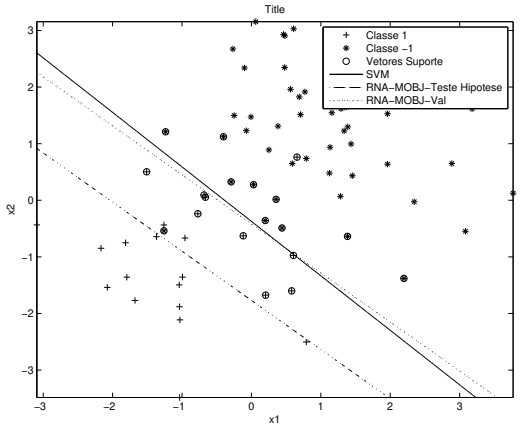


Figura 5.18: Curvas de separação: Decisão por Validação e por Hipótese, com p = 0.02.

classificador privilegiar os exemplos da classe -1. Na Figura 5.16 também foi evidenciado que a escolha de uma probabilidade *a priori* de erros nos rótulos do conjunto de dados era inferior ao que havia sido inserido nos dados. Portanto, o classificador com probabilidade *a priori* $p = 0.05$ apresentou-se excessivamente ajustado aos dados, não obtendo uma solução de alta generalização.

5.2 Problema de Duas Distribuições em forma de Lua

Esta base de dados foi obtida à partir de duas funções de distribuição bidimensionais em forma de meia lua e com um nível de ruído desconhecido. Essa base apresenta não linearidade em sua curva de decisão, porém, não apresenta sobreposição da região limite entre as classes.

A base possui 400 exemplos divididos igualmente em duas classes e será usada para analisar o aprendizado de uma RNA com o método MOBJ com decisão probabilística e por erro de validação. Além disso, os resultados obtidos serão confrontados com uma SVM treinada com a mesma base de dados. O conjunto de soluções Pareto-ótimas foi gerado por 21 redes MLP com 10 neurônios na camada oculta.

As análises seguintes referem-se aos experimentos realizados para destacar as características do decisor probabilístico.

5.2.1 Análise da Sobreposição das Classes

Aqui os experimentos realizados servem para destacar a influência da sobreposição espacial dos dados na obtenção de uma curva de separação entre as classes. A condição de sobreposição dos exemplo do conjunto de dados serve para dificultar o processo de aprendizagem e, consequentemente, avaliar a capacidade do decisor em escolher a solução Pareto-ótima que melhor represente os dados problema.

É possível verificar um razoável grau de sobreposição entre os exemplos de cada classe. O aprendizado deve obter uma curva de separação entre os exemplos, sem que faça um ajuste excessivo dos parâmetros ajustáveis do modelo, impedindo que o modelo seja "forçado" a considerar corretamente todos os rótulos dos exemplos pertencentes à região de fronteira entre as classes. Nessa região encontram-se os exemplos com maior probabilidade de apresentarem algum ruído que o rotularam erroneamente.

Considerando, o conhecimento *a priori* que um especialista pode possuir em problemas dessa natureza, o mesmo pode, então, usar dessa informação para controlar a complexidade da solução escolhida pelo decisor ao informar o grau de confiança no processo de coleta dos dados.

Os experimentos realizados destacam as curvas de separação dos decisores por erro de validação, o probabilístico, da SVM e apresenta, também, a *Curva L* do conjunto de soluções (hipóteses) geradas. Para o decisor baseado no teste estatístico de hipóteses foram realizadas variações paramétricas do valor p, referente a informação *a priori* usada no processo de aprendizagem.

Os experimentos realizados usaram uma rede MLP com estrutura definida por 20 neurônios da camada oculta, função de transferência tangente hiperbólica para as camadas oculta e de saída. O algoritmo multiobjetivo foi parametrizado de tal maneira a iniciar a geração de soluções com norma mínima $||w_{min}|| = 0.5$, norma máxima $||w_{max}|| = 15$ e incremento $\Delta||w|| = 0.5$. Cada solução era gerada com o algoritmo Levenberg-Marquardt determinando os parâmetros ajustáveis da rede a fim de obter a rede de mínimo erro com a restrição de norma especificada. Desse modo, o conjunto Pareto-ótimo foi constituído de 31 soluções. O conjunto de dados usado continha 400 exemplos, dos quais 70% foi usado para aprendizado e os 30% eram disponíveis para o teste de acurácia do classificador. Dos 70% restantes, ao usar o treinamento multiobjetivo com decisão por erro de validação, o conjunto foi novamente fracionado em 70% para a etapa de treinamento e os 30% restantes foram usados para a etapa de decisão. No treinamento multiobjetivo com decisão probabilística todos os exemplos disponíveis para o processo de aprendizagem foram usados na etapa de treinamento. Os experimentos foram realizados usando a seguinte parametrização do valor p de informação *a priori*: 0.05, 0.1, 0.15, 0.2 e 0.3.

Para compreensão do conceito de acurária, métrica comumente usada em aprendizagem de máquina, ela se refere ao quão frequente o classificador está correto. E pode ser representada pela Equação 5.1 , sem perda de generalidade, para classificadores binários.

$$acc = \frac{VP + VN}{total} \tag{5.1}$$

onde VP é a taxa de verdadeiros positivos, VN é a taxa de verdadeiros negativos e *total*

é o número total de dados classificados.

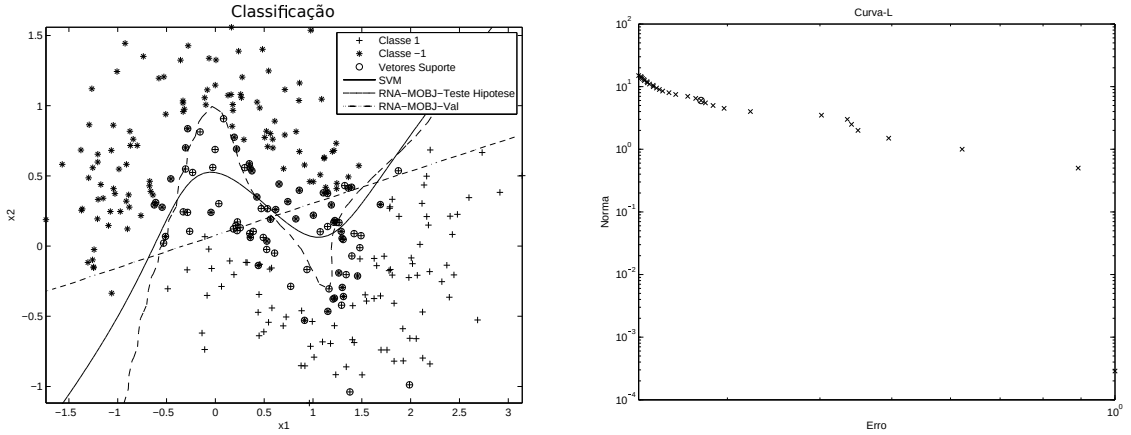


Figura 5.19: Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.05$.

Avaliando a Figura 5.19 verificou-se que o valor $p = 0.05$ indicava a tendência por um modelo excessivamente complexo. Além disso, o decisor por erro devalidação, devido a natureza aleatória de amostragem, resultou na escolha de um modelo sub-dimensionado. A SVM apresentou o resultado consistente esperado. Ainda na Figura 5.19 a *Curva L* obtida é pouco esclarecedora e não permitiu que a região de máxima curvatura, ou seja, o *corner*, fosse facilmente identificado.

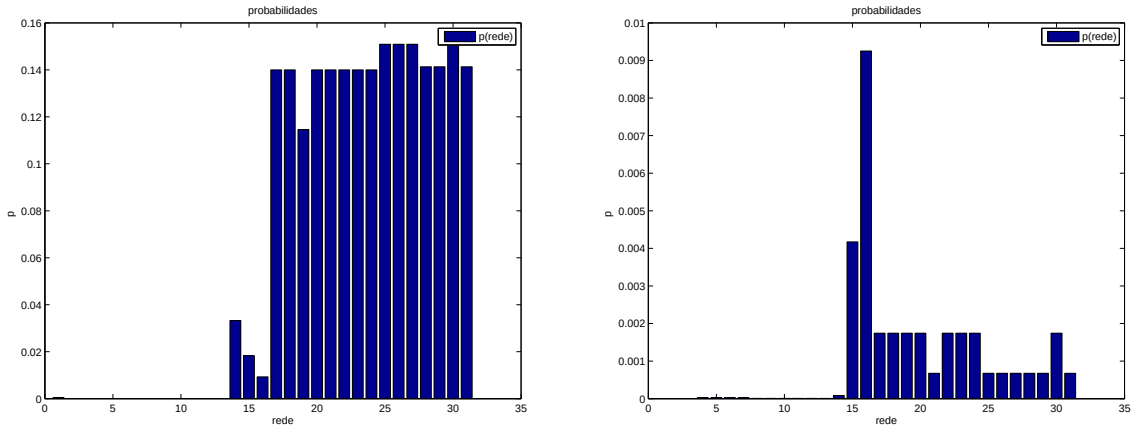


Figura 5.20: Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.05$.

Em oposição aos resultados obtidos usando o valor p significativamente baixo, a Figura 5.21 exibe o resultado do aprendizado multiobjetivo com $p = 0.3$. Assim foi possível verificar a influência de valores p extremistas, gerando, respectivamente, modelos super-dimensionados e sub-dimensionados.

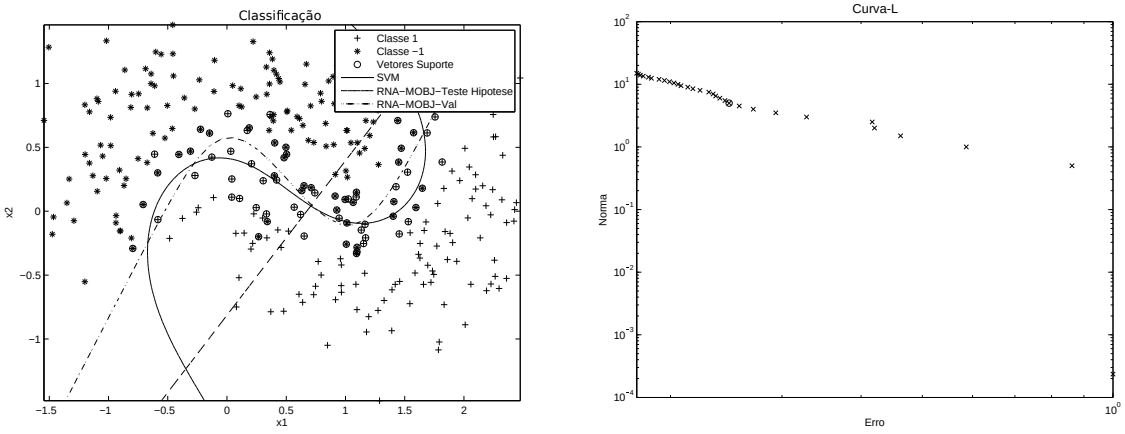


Figura 5.21: Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.30$.

Agora, avaliando a Figura 5.21 foi possível identificar o efeito de um valor elevado para a probabilidade de erros nos dados, destacando uma solução excessivamente suave da curva de separação obtida pelo decisor probabilístico. As probabilidades das redes representarem estes dados estão indicadas na Figura 5.22 onde o pico da distribuição binomial está entre as redes de menor complexidade.

No entanto, a condição o conhecimento do valor adequado de p define a qualidade do modelo obtido. Assim, o valor $p = 0.1$, indicando o nível de classificações errôneas presentes em cada classe do conjunto de dados, indica o valor dessa informação quando conhecida com acurácia. Dentre o modelos do conjunto Pareto-ótimo, o modelo mais provável, segundo a distribuição binomial gerada, está representado na Figura 5.23.

Ainda na Figura 5.23 a *Curva L* representada indicava uma região de *corner* um pouco mais evidente que coincidiu com a localização da solução escolhida pelo decisor probabilístico.

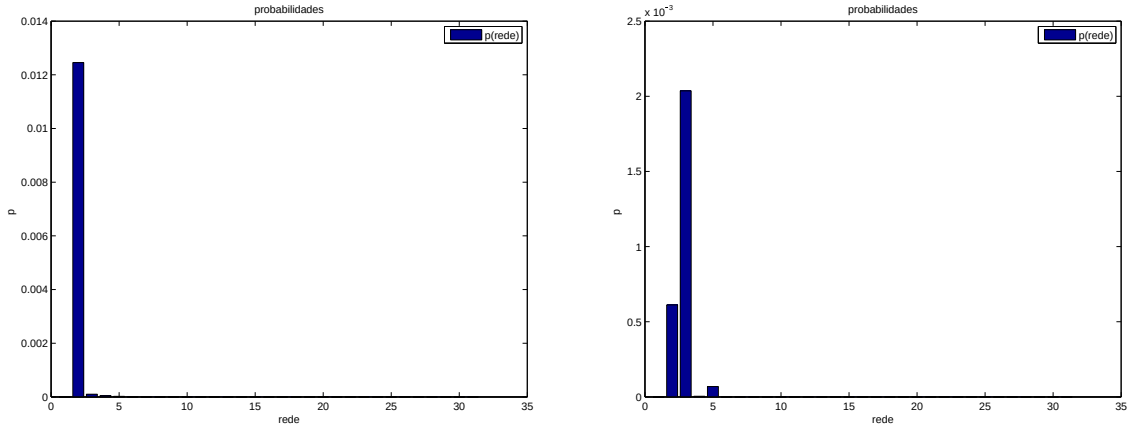


Figura 5.22: Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.30$.

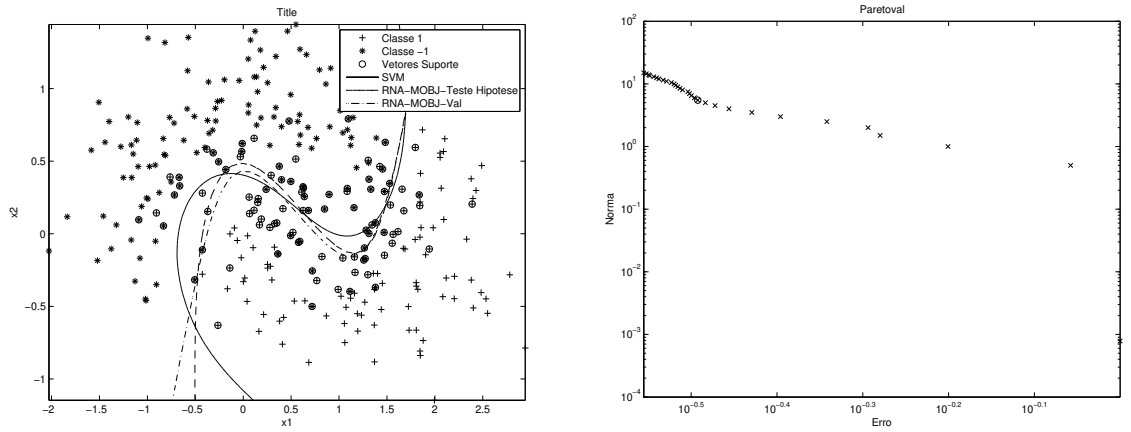


Figura 5.23: Curvas de Separação e *Curva L* gerada com dados do treinamento sobrepostos e com decisor probabilístico com $p = 0.10$.

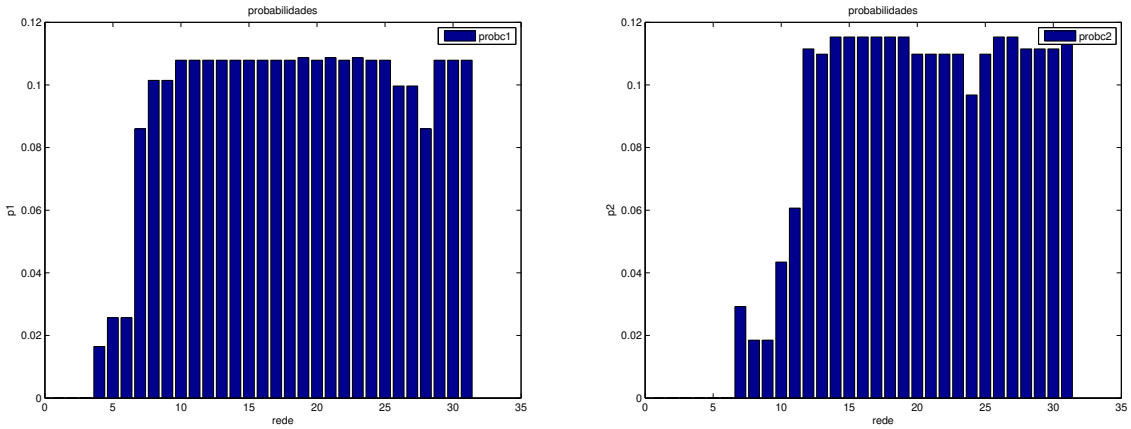


Figura 5.24: Distribuições Binomiais com indicação da rede mais provável de cada classe usando $p = 0.10$.

5.3 Problema de Diagnóstico do Coração

O objetivo dos experimentos com esta base era descobrir as classes de pacientes com problemas cardíacos e sadios. A base possuia 13 características e a classe indicando o diagnóstico do paciente. Eram 270 exemplos de pacientes. A base foi apresentada a uma RNA com o método MOBJ sob duas diferentes estratégias de decisão e também apresentada a uma uma SVM de kernel polinomial. Além disso, um outro objetivo com estes experimentos era mostrar que em problemas complexos como estes poderíamos ter soluções melhores diante do uso eficiente dos dados disponíveis e, principalmente, do uso da informação *a priori* correta na seleção do modelo.

Como tratava-se de uma base multi-dimensional, não havia como mostrar a superfície de separação obtida em cada estratégia e, portanto, os resultados foram obtidos por meio da taxa de acertos de diagnósticos de cada método em um conjunto de teste. Este conjunto correspondia a exemplos não utilizados para treinamento e nem validação.

O método MOBJ foi utilizado em uma RNA MLP com 20 neurônios na camada oculta, com funções de transferência do tipo tangente hiperbólica nas camadas ocultas e de saída, e com o algoritmo de otimização Levenberg-Marquardt com limite máximo de 500 épocas. Foram geradas 16 soluções com variação de norma dos pesos igual a 1. Para cada norma, foi obtida a rede com o mínimo erro de treinamento. Neste conjunto, foi tomada a decisão por erro de validação e probabilística.

5.3.1 A Base de Dados

A base de dados usada pertence ao repositório de dados público sobre problemas de aprendizado de máquina da *University of California at Irvine* (Asuncion and Newman, 2007). A base correspondente ao problema do diagnóstico de problemas do coração é uma das mais difundidas desse repositório e, normalmente, é uma base de difícil diagnóstico.

Originalmente, a base de dados contém 76 atributos para cada indivíduo. Contudo, devido ao grande número de exemplos incompletos na maioria dos atributos, todos os experimentos publicados faziam referência à utilização de somente 13 destes (mais a classe ou atributo-meta), os quais foram listados a seguir:

- **Idade** (atributo real entre 29 e 77 anos);
- **Sexo** (atributo binário, sendo masculinos representados por 1 e femininos por 0);
- **Tipo de Dor Toráxica** (atributo nominal, sendo 4 possíveis tipos de dores:);
    1. Angina Típica.
    2. Angina Atípica.
    3. Sem Dor Anginal.
    4. Assintomático.
- **Pressão Arterial em Repouso** (atributo real medido em mmHg);
- **Colesterol no Soro** (atributo real medido em mg/dl);
- **Concentração de Açúcar no Sangue > 120 mg/dl** (atributo binário, sendo verdadeiro representado por 1 e falso por 0);
- **Resultado da Eletrocardiografia em Repouso** (atributo nominal, sendo 4 possíveis valores:);
    1. Normal.
    2. Com onda ST-T anormal.
    3. Mostrando provável (ou definida) hipertrofia do ventrículo esquerdo.
- **Frequência Cardíaca Máxima Atingida** (atributo real medido em bpm);
- **Angina Induzida por Exercício** (atributo binário, sendo SIM representado por 1 e NÃO por 0);
- **Depressão ST induzida por exercício relativamente sossegado** (atributo real)
- **Inclinação da extremidade do segmento ST no exercício** (atributo ordinal, sendo 3 possíveis valores:)

1. Inclinado para Cima.
2. Plano.
3. Inclinado para Baixo.

- **Número de vasos coloridos pela fluoroscopia** (atributo real entre 0 e 3);
- **Talassemias** (atributo nominal, sendo 3 possíveis valores:)
  - **Valor 3**: Normal.
  - **Valor 6**: Defeito Fixo (Irreparável).
  - **Valor 7**: Defeito Reversível (Reparável).
- **Diagnóstico** (atributo binário, sendo os indivíduos com menos de 50% de estreitamento do diâmetro do vaso sanguíneo representados por -1, e os indivíduos com 50% ou mais de estreitamento do diâmetro do vaso sanguíneo representados por 1)

Durante os experimentos não foi realizada nenhuma atividade de pré-processamento como, por exemplo, a seleção de exemplos mais relevantes. O objetivo principal nos experimentos era a análise das estratégias de decisões do método MOBJ sob o conjunto de dados.

5.3.2 Experimentos

O treinamento da rede *MultiLayer Perceptron* foi realizado com todo o conjunto de dados, uma vez que a abordagem de decisão probabilística dispensa a separação das amostras. O conjunto de teste foi usado para avaliar o desempenho final da rede escolhida. Assim sendo, foram geradas as aproximações das distribuições binomiais indicativas das redes mais prováveis para cada classe do problema. A Figura 5.25 representa as probabilidades das soluções para as classes dos doentes cardíacos e não-cardíacos, respectivamente.

O conjunto de treinamento usado nos experimentos corresponde à 70% do conjunto de dados e o restante é o conjunto de teste. Resumindo, eram 189 dados para treinamento e 81 dados para teste. Os exemplos de treinamento foram selecionados aleatoriamente, com igual probabilidade para todos. Os exemplos de teste foram separados para estimar o erro de generalização. O decisor probabilístico trabalhou com a probabilidade de erro dos rótulos em 5% uma vez que este nível de significância estatística é, para muitas áreas de estudo, estatisticamente relevante na realização de testes de hipóteses (Krzywinski and Altman, 2013).

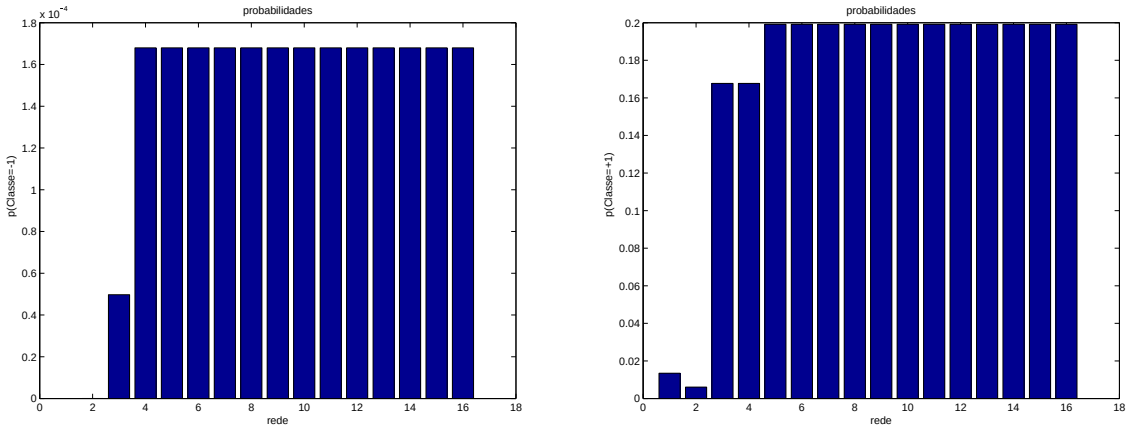


Figura 5.25: Aproximações das Distribuições Binomiais indicando as probabilidades de cada rede para a classe dos cardíacos e não-cardíacos.

A Figura 5.25 mostra a aproximação da distribuição binomial das redes mais prováveis para a classe dos pacientes cardíacos e não-cardíacos, respectivamente. Para ambas situações, entre duas ou mais hipóteses com a mesma probabilidade era feita a escolha pela hipótese mais simples. Dessa forma era possível indicar a solução mais provável em cada classe. Como não foi definido um critério para se escolher qual hipótese representaria a mais provável para ambas as classes simultaneamente, foi usada uma estratégia baseada na média das duas.

A Figura 5.26 mostra o conjunto de soluções Pareto-ótimas e a melhor solução, indicada por um círculo (o), para as decisões probabilísticas e por erro validação, respectivamente.

A Figura 5.27 mostra as *Curvas L* discretas com a solução escolhida pelo decisor, indicada por um círculo (o), para as decisões probabilísticas e por erro de validação,

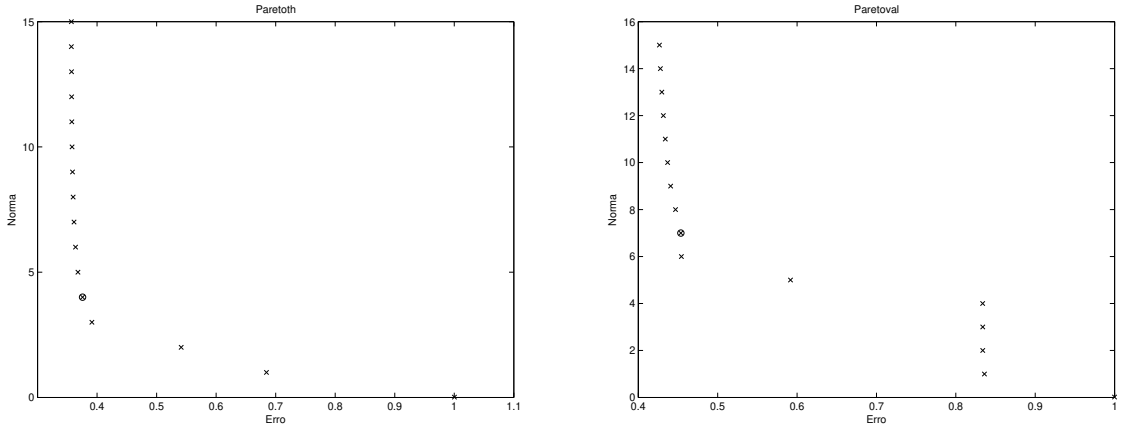


Figura 5.26: Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.05) e por erro de validação (70% dos exemplos de treinamento)

respectivamente.

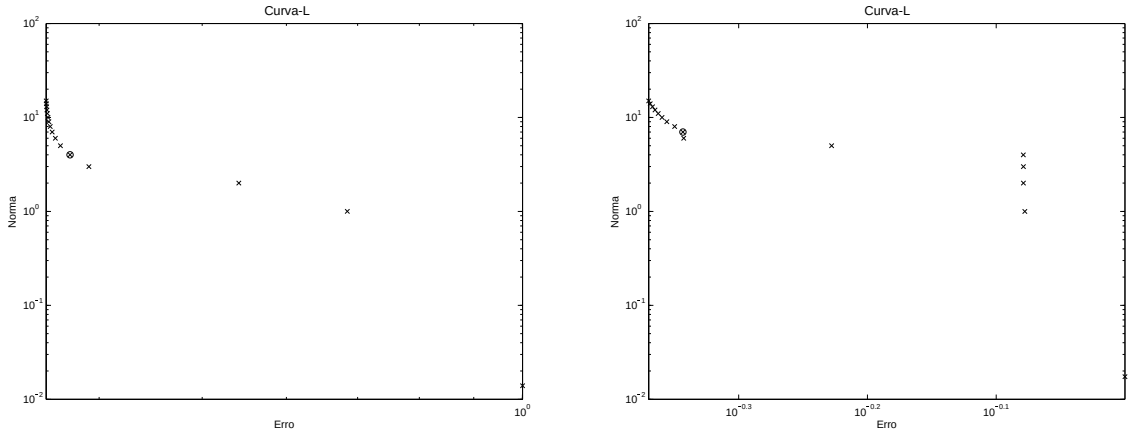


Figura 5.27: A *Curva L* de redes obtidas com estratégia de decisão probabilísticas (p=0.05) e por erro de validação (70% dos exemplos de treinamento)

De acordo com o critério de decisão da *Curva L*, a região de máxima curvatura indicava a solução que melhor equilibrava os efeitos de polarização e de variância. Portanto, de acordo com as duas aproximações da *Curva L*, os decisores indicaram uma solução localizada próxima à região de máxima curvatura. Isto indica que, provavelmente, ambos os critérios de decisão foram adequados para o conjunto de soluções Pareto-ótimas.

5.3.3 Acurácia dos Classificadores

O resultado das simulações executadas com a SVM e com as duas redes MLP foram comparados através da taxa de acertos sob o conjunto de teste. A SVM apresentou um desempenho constante e superior às demais redes de 88,89% de acerto. A rede MLP com decisão por hipótese apresentou um desempenho de 85,19% de acerto, superior à rede com decisão por validação com 75,31% de acerto.

A baixa acurácia do classificador obtido por decisão com um conjunto de validação é facilmente justificada uma vez que a base de validação é selecionada aleatoriamente a cada experimento. Isto permite que, eventualmente, o conjunto de validação usado não represente a uma aproximação da distribuição real das classes. Nestes casos, temos uma decisão polarizada. A matriz de confusão representada na Tabela 5.3, mostra para qual classe a decisão ficou polarizada.

Matriz de Confusão

A matriz de confusão foi usada para permitir uma análise mais detalhadas dos experimentos das redes MLP e da SVM, indicando os dois possíveis erros de classificação: os falsos positivos e os falsos negativos. A identificação da taxa de verdadeiros positivos (a sensibilidade) e a taxa de verdadeiros positivos (a especificidade) pode tornar mais clara a diferença entre cada modelo, especialmente na avaliação de métodos para diagnósticos médicos. Obviamente, a minimização de ambas configuraria o caso perfeito. No entanto, pode-se compreender que um modelo capaz de minimizar a taxa de falsos negativos pode ser preferido a um modelo que minimize a taxa de falsos positivos. Em um cenário real, os casos positivos sempre passam por uma nova avaliação para confirmação do diagnóstico. Em contra-partida, os casos negativos não costumam ser verificados porque, geralmente, são estatisticamente mais garantidos. Por isso a acurácia geral de um modelo não diz tudo sobre sua capacidade, por isso a matriz de confusão é uma importante ferramenta para análise das medidas de sensibilidade e especificidade dos métodos.

A distribuição das classes no conjunto de teste seguiu a proporção de 55,56% de pacientes sadios (45 pacientes sadios) e, consequentemente, de 44,44% de pacientes com doença do coração (36 pacientes). A Tabela 5.1 resume os acertos do classificador SVM. A Tabela 5.2 resume os acertos do classificador gerado pela rede com decisão probabilística, enquanto a Tabela 5.3 resume os acertos do classificador gerado pela rede com validação.

As matrizes de confusão montadas neste trabalho apresentam detalhadamente os tipos de acertos e erros de cada um dos classificadores gerados com a base de dados. A Tabela 5.1 mostra o desempenho da Máquina de Vetores de Suporte e as Tabelas 5.2, 5.3 mostram o desempenho das redes MLP no conjunto de teste.

Tabela 5.1: Matriz de Confusão do classificador SVM.

Previsto/Dado	Sadio	Não-sadio
Sadio	39	03
Não-sadio	06	33

Os resultados obtidos pela SVM mostram elevada acurácia na previsão do diagnóstico médico. A análise dos erros mostrou que a SVM apresentou a maior parte dos erros no diagnóstico de falsos-negativos. Isto apontava que quando a SVM errava uma classificação, o erro mais provável era o diagnóstico positivo (paciente cardíaco) quando o paciente não apresentava a doença. Este erro é, notavelmente, menos crítico. Nessas situações, um resultado positivo no exame pode ser reavaliado por outro tipo de exame para validar o resultado anterior.

Os resultados obtidos pela rede MLP com decisor probabilístico mostraram que o erro mais provável de ocorrer durante um diagnóstico era o diagnóstico negativo (paciente sadio) quando o paciente apresentava a doença. Este erro é crítico e deveria ser evitado. Mesmo a rede neural apresentando um desempenho superior aos 90% de acertos nos casos de diagnósticos negativos, os erros encontrados pelo classificador correspondiam

Tabela 5.2: Matriz de Confusão do classificador obtido por decisão probabilística.

Previsto/Dado	Sadio	Não-sadio
Sadio	41	06
Não-sadio	04	30

ao tipo mais severo. Portanto é sugerido que qualquer diagnóstico negativo seja testado novamente antes de se afirmar que o paciente está realmente saudável.

Tabela 5.3: Matriz de Confusão do classificador obtido por erro nos dados de validação.

Previsto/Dado	Sadio	Não-sadio
Sadio	30	05
Não-sadio	15	31

Os resultados observados na Tabela 5.3 mostraram que o decisor por erro de validação escolheu uma rede com desempenho abaixo de 80% de acertos. Isso gerou um percetual bastante elevado de falsos-positivos, 15 diagnósticos. Enquanto isso, o número de verdadeiros-negativo foi de 30 diagnósticos, gerando mais gastos com novos exames para comprovarem a eficácia do diagnóstico. A resposta para o fraco desempenho da rede pode ser a insuficiente densidade da amostra de treinamento, uma vez que foi necessário separar uma fração desse conjunto para fazer a seleção do modelo. Ao mesmo tempo, a decisão baseada no conjunto de validação torna a escolha por um modelo com mínimo erro de generalização dependente da distribuição das classes.

Avaliação da Variação do Valor *a priori*

A variação do valor *a priori* p na estratégia de decisão é um fator preponderante na construção de uma solução com alta capacidade de generalização. Para isso foram realizadas duas variações desse parâmetro com o objetivo de estudar esses dois cenários. Os valores p usados foram: 0.01 e 0.4. Para cada nova parametrização de p o decisor destacou a rede mais provável diante dos dados de cada classe. As Figuras 5.28 e 5.29 destacam as redes mais prováveis para cada classe do conjunto de dados fornecido. A região entre as soluções destacadas corresponde à região de interesse para o problema de aprendizado.

Os resultados obtidos destacaram a influência desse valor em cenários completamente distintos onde, no primeiro caso, o valor baixo de p ($p = 0.01$) indicava a elevada confiança na qualidade dos dados. No segundo caso, o valor elevado de p ($p = 0.4$) indicava a baixa confiança na qualidade dos dados. Nesse dois cenários houve mudanças profundas na construção do classificador. A Figuras 5.28 e 5.29 mostram as curvas Pareto-ótimas obtidas, em escala logarítmica, para $p = 0.01$ e $p = 0.4$, respectivamente.

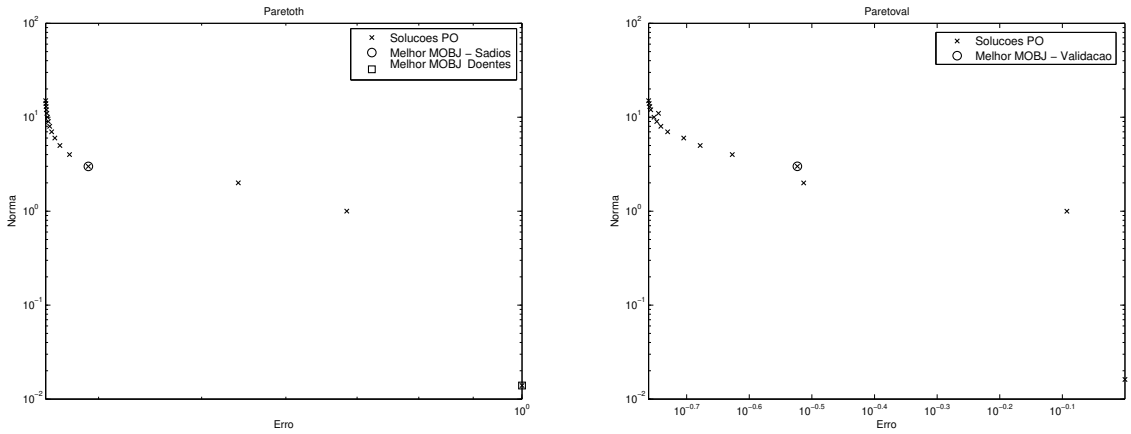


Figura 5.28: Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.01) e por erro validação (70% dos exemplos de treinamento)

A *Curva L* da Figura 5.28 mostra as soluções mais indicadas para cada classe do conjunto de dados com o valor $p = 0.01$. Por outro lado, a *Curva L* da Figura 5.29 mostra as soluções mais indicadas para cada classe do conjunto de dados com o valor $p = 0.4$.

A acurácia do classificador nesses dois casos foi comprometida por dois fatores: o erro causado pelo ajuste excessivo aos exemplos devido à elevada confiança no conjunto de dados e o erro causado pelo ajuste insuficiente aos exemplos devido à baixa confiança no conjunto de dados. Dessa forma, a presença do conhecimento do especialista do domínio era fundamental na indicação da região de soluções que melhor representava o problema de aprendizado por meio do conjunto de dados disponível.

A referência usada para avaliar a acurácia dos classificadores obtidos pelo método

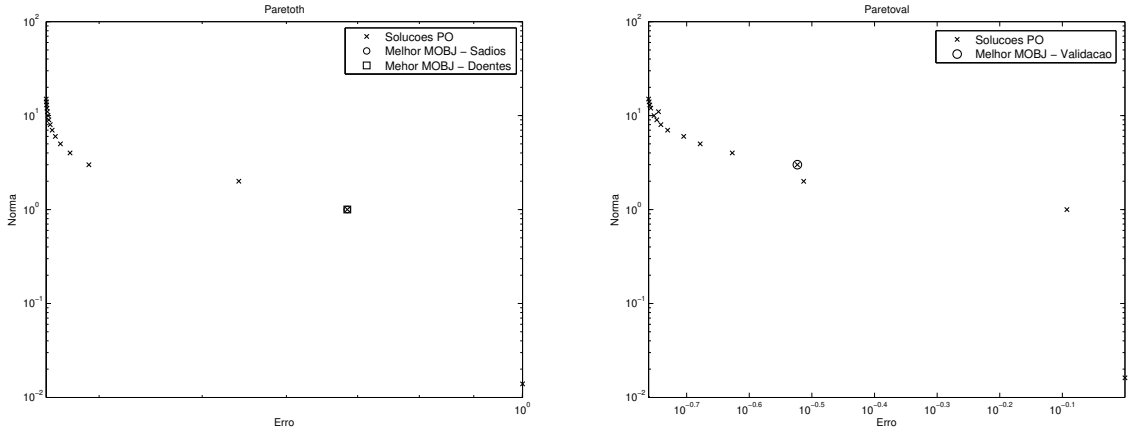


Figura 5.29: Conjunto Pareto de redes obtidas com estratégia de decisão probabilística (p=0.4) e por erro de validação (70% dos exemplos de treinamento)

MOBJ com decisão probabilística foi baseada no desempenho com os dados de teste, conforme a Figura 5.30. Assim foi possível verificar que o desempenho das redes MLP com método MOBJ possuíam desempenho numericamente idênticos a partir de uma determinada faixa de valores para a norma $||w||$.

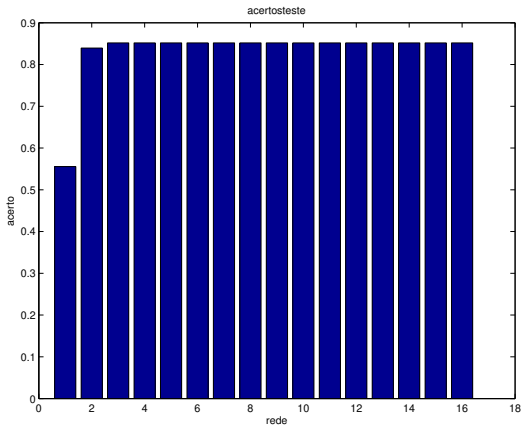


Figura 5.30: Desempenho das redes MOBJ no conjunto de dados de teste.

Finalizando essa análise da variação do valor p para o decisor probabilístico, a Tabela 5.4 mostra a rede mais provável para representar os dados de cada classe mediante o valor definido para p no decisor. Os casos em que as decisões mais se distanciavam eram as circunstâncias em que o valor p assumia um valor muito baixo, como, por exemplo, $p = 0.01$ e $p = 0.02$. Nesses dois exemplos citados, as redes escolhidas eram bem diferentes uma da outra. Isto ocorria porque o decisor definia qual rede era a melhor para acertar todos os casos de cada classe individualmente. Portanto, a rede com norma baixa era escolhida para representar uma classe enquanto a rede com norma elevada era escolhida

para representar a outra classe.

Os outros resultados apresentados pela Tabela 5.4 mostraram que a decisão confirmou sua tendência em chegar num consenso sobre qual das redes MLP era a mais provável, com algumas pequenas variações, mas pouco representativas.

Tabela 5.4: Sensibilidade do decisor probabilístico à variações do valor p .

Valor p	Nºda rede - Classe Sadios	Nºda rede - Classe Doentes
0.01	4	1
0.02	4	1
0.03	4	5
0.04	4	5
0.05	4	3
0.06	4	3
0.07	4	3
0.08	4	3
0.09	4	3
0.10	4	2
0.15	4	2
0.20	3	2
0.25	2	2
0.30	2	2
0.40	2	2
0.50	2	2

5.4 Comentários Finais

Neste capítulo foram apresentados os resultados de experimentos realizados com o treinamento de SVMs e de rede MLPs com o método MOBJ usando duas estratégias de decisão em alguns problemas de dados sintéticos e um problema real de classificação de padrões. O objetivo principal era mostrar a eficiência do método MOBJ com a nova estratégia de decisão baseada em um teste estatístico de hipóteses. Foi mostrado que, quando disponível, uma informação prévia da probabilidade de classificações errôneas nos dados poderia auxiliar a tomada de decisão no conjunto de soluções Pareto-ótimas. Embora não seja sempre que essa informação *a priori* esteja disponível, na sua presença é possível direcionar a busca do decisor. Sugere-se que quando não houver nenhuma informação prévia da confiabilidade dos dados, que seja utilizado o decisor baseado no erro de validação, uma vez que o mesmo escolhe uma solução razoável no conjunto Pareto-ótimo. No entanto, na presença da informação prévia sobre a qualidade dos dados disponíveis, o processo de tomada de decisão no conjunto Pareto-ótimo pode tornar-se mais eficiente. Isto ocorre porque o processo de decisão utiliza agora um princípio estatístico para evidenciar qual das hipóteses (soluções) não-dominadas é a mais plausível de representar a função geradora dos dados. Além disso, por consequência, o conjunto de treinamento tornar-se-á mais representativo uma vez que os dados anteriormente usados para compor o conjunto de validação poderão agora ser usados para compor um conjunto de treinamento mais numeroso.

Capítulo 6

Estratégia de Decisão em Regressão

Em estatística, a autocorrelação é uma medida que informa o grau de dependência entre os valores assumidos por uma variável aleatória em diferentes instantes. Por exemplo, o quanto a existência de valor mais alto condiciona valores também altos de seus vizinhos.

Existem várias interpretações físicas da autocorrelação, e mesmo várias definições. Segundo a definição da estatística, o valor da autocorrelação está entre 1 (correlação perfeita) e -1, o que significa anti-correlação perfeita. O valor 0 significa total ausência de correlação (Peebles, 1993; Box and Jenkins, 1990).

As amostras em um problema de aprendizagem de máquina podem apresentar algum ruído e a hipótese considerada para usar a informação da autocorrelação do resíduo entre a função aproximada e os dados é a de que o ruído corresponde a uma variável aleatória de distribuição normal com média zero, uma variância $\sigma^2 > 0$ e apresenta uma autocorrelação nula, ou seja, o ruído é uma variável com total ausência de correlação. Sob esta hipótese, foi construída uma estratégia de decisão para os problemas de regressão cuja busca pela solução no conjunto Pareto-ótimo encontre a solução de menor autocorrelação do resíduo gerado.

6.1 A Regra de Decisão por Autocorrelação

Examina-se agora o problema de regressão, no qual uma rede neural deve modelar a função $f(x)$ dada por:

$$y = f(x) + \xi \tag{6.1}$$

Para a síntese da rede neural, é possível acessar pares entrada-saída $(x_i, y_i) = (x_i, f(x_i) + \xi_i)$, intrinsicamente contaminados com um ruído aleatório ξ. Assume-se aqui que esse ruído tenha média zero e seja não autocorrelacionado, ou seja:

$$\mathbb{E}(\xi_i) = 0 \ \forall \ i \qquad \mathbb{E}(\xi_i \cdot \xi_j) = 0 \ \forall \ i \neq j \tag{6.2}$$

Também se assume que o ruído ξ não seja correlacionado com a função $f(x)$, ou seja:

$$\mathbb{E}(\xi_i \cdot f(x_j)) = 0 \ \forall \ i, j \tag{6.3}$$

A rede neural é representada pela função:

$$\hat{y} = \hat{f}(x, w) \tag{6.4}$$

O treinamento da rede neural é realizado por uma técnica multiobjetivo que conduz à obtenção de um conjunto W^* de vetores de pesos que são Pareto-ótimos, no sentido discutido nos capítulos anteriores desta tese. Neste ponto, passa-se à discussão sobre a escolha de um vetor $w^* \in W^*$ que produza a melhor aproximação para a função. Esta tese propõe a utilização da *regra de mínima autocorrelação do resíduo*, que escolhe a solução Pareto-ótima com o vetor de pesos ótimo w^* dada pela Equação:

$$w^* = \arg \min_{w \in W^*} \mathcal{A}(r(w)) \tag{6.5}$$

Nessa expressão, $r(w)$ representa a sequência de resíduos verificados entre o modelo correspondente à rede neural com vetor de pesos w e o conjunto de dados de treinamento contendo p pares (x_i, y_i):

$$r(w) = [r_i(w)]_{i=1}^{p} = [y_i - \hat{f}(x_i, w)]_{i=1}^{p} = [f(x_i) + \xi_i - \hat{f}(x_i, w)]_{i=1}^{p} \tag{6.6}$$

Para se definir $\mathcal{A}(r(w))$, preliminarmente definimos a sequência $\tilde{r}(w)$, que corresponde à sequência $r(w)$ acrescida de $p-1$ zeros no início da sequência:

$$\tilde{r}(w) = [0]_{i=1}^{p-1} \oplus [r_i(w)]_{i=1}^{p} \tag{6.7}$$

sendo que nesta expressão o operador $\oplus$ significa a concatenação de sequências. Convenciona-se que a sequência $\tilde{r}$ seja indexada de forma que seu primeiro elemento tenha o índice $1-p$ e seu último elemento tenha o índice p. O funcional $\mathcal{A}(r(w))$ agora é definido como:

$$\mathcal{A}(r(w)) = \sum_{\ell=1}^{p-1} \left| \sum_{i=1}^{p} \tilde{r}_i(w) \cdot \tilde{r}_{i-\ell}(w) \right| \tag{6.8}$$

Os termos $\tilde{r}_i \cdot \tilde{r}_{i-\ell}$ não nulos têm a forma:

$$\tilde{r}_i(w) \cdot \tilde{r}_{i-\ell}(w) = (f(x_i) + \xi_i - \hat{f}(x_i, w)) \cdot (f(x_{i-\ell}) + \xi_{i-\ell} - \hat{f}(x_{i-\ell}, w)) \tag{6.9}$$

Rearranjando esses termos obtém-se:

$$\begin{aligned} \tilde{r}_i(w) \cdot \tilde{r}_{i-\ell}(w) \quad = \quad & f(x_i) \cdot f(x_{i-\ell}) + \xi_i \cdot \xi_{i-\ell} + \hat{f}(x_i, w) \cdot \hat{f}(x_{i-\ell}, w) + \\ & f(x_i) \cdot \xi_{i-\ell} - f(x_i) \cdot \hat{f}(x_{i-\ell}, w) + \xi_i \cdot f(x_{i-\ell}) - \\ & \xi_i \cdot \hat{f}(x_{i-\ell}, w) - \hat{f}(x_i, w) \cdot f(x_{i-\ell}) - \hat{f}(x_i, w) \cdot \xi_{i-\ell} \end{aligned} \tag{6.10}$$

Utilizando as premissas de que o ruído ξ_i não tenha autocorrelação e que não seja correlacionado com a função $f(x)$, obtém-se:

$$\mathbb{E}(\xi_i \cdot \xi_{i-\ell}) = 0 \qquad \mathbb{E}(\xi_i \cdot f(x_{i-\ell})) = \mathbb{E}(f(x_i) \cdot \xi_{i-\ell}) = 0 \tag{6.11}$$

A expressão da esperança de $\tilde{r}_i(w) \cdot \tilde{r}_{i-\ell}(w)$ fica então:

$$\begin{aligned} \mathbb{E}(\tilde{r}_i(w) \cdot \tilde{r}_{i-\ell}(w)) \quad = \quad & \mathbb{E}(f(x_i) \cdot f(x_{i-\ell}) - f(x_i) \cdot \hat{f}(x_{i-\ell}, w)) + \\ & \mathbb{E}(\hat{f}(x_i, w) \cdot \hat{f}(x_{i-\ell}, w) - \hat{f}(x_i, w) \cdot f(x_{i-\ell})) - \\ & \mathbb{E}(\xi_i \cdot \hat{f}(x_{i-\ell}, w) - \hat{f}(x_i, w) \cdot \xi_{i-\ell}) \end{aligned} \tag{6.12}$$

Se existir algum $w^* \in W^*$ tal que $f(x) \approx \hat{f}(x, w^*)$, nesse caso serão válidas as expressões:

$$\begin{aligned} &\mathbb{E}(f(x_i) \cdot f(x_{i-\ell}) - f(x_i) \cdot \hat{f}(x_{i-\ell}, w^*)) \approx 0 \\ &\mathbb{E}(\hat{f}(x_i, w^*) \cdot \hat{f}(x_{i-\ell}, w^*) - \hat{f}(x_i, w^*) \cdot f(x_{i-\ell})) \approx 0 \end{aligned} \tag{6.13}$$

Nessa situação, também serão válidas:

$$\begin{aligned} &\mathbb{E}(\xi_i \cdot \hat{f}(x_{i-\ell}, w^*)) \approx \mathbb{E}(\xi_i \cdot f(x_{i-\ell})) = 0 \\ &\mathbb{E}(\hat{f}(x_i, w^*) \cdot \xi_{i-\ell}) \approx \mathbb{E}(f(x_i) \cdot \xi_{i-\ell}) = 0 \end{aligned} \tag{6.14}$$

Desta forma, caso exista tal $w^* \in W^*$, chega-se à conclusão de que

$$\mathbb{E}(\tilde{r}_i(w^*) \cdot \tilde{r}_{i-\ell}(w^*)) \approx 0, \tag{6.15}$$

o que conduz a:

$$\mathbb{E}(\mathcal{A}(r(w^*))) \approx 0 \tag{6.16}$$

Da definição de $\mathcal{A}(r(w))$, observa-se que

$$\mathbb{E}(\mathcal{A}(r(w))) \geq 0 \ \forall \ w \tag{6.17}$$

o que significa que a determinação de w^* pode ser feita pela minimização de $\mathcal{A}(r(w))$ em $w \in W^*$.

6.2 Comentários Finais

Neste capítulo foi apresentada uma estratégia de decisão por soluções não-dominadas para os problemas de regressão, baseada no princípio da mínima autocorrelação do resíduo. O resíduo estimado pela diferença entre as saídas conhecidas dos dados de treinamento e as saídas da função aproximada deve apresentar as características de uma variável aleatória de distribuição normal. Então, sob esta abordagem, o resíduo estimado pela melhor solução do conjunto Pareto-ótimo deverá ser constituído, praticamente, pelo ruído. Portanto, este decisor é orientado a escolher a solução cuja medida da autocorrela-

ção do resíduo aproxime-se de zero, indicando um bom ajuste da rede MLP ao conjunto de dados disponível. O princípio dessa estratégia de decisão é conceitualmente válido com a suposição de ruído não correlacionado e, inclusive, com ruídos correlacionados, desde que não sejam excessivamente correlacionados. O capítulo de simulação em problemas de regressão destacará a eficiência do decisor em alguns problemas com ruído não correlacionado e correlacionado para ilustrar sua base teórica.

Capítulo 7

Simulações da Decisão em Regressão

Neste capítulo são apresentados os resultados da aplicação da estratégia de decisão por mínima autocorrelação em modelos obtidos pelo método MOBJ através de experimentos computacionais em problemas de regressão.

O objetivo deste capítulo é destacar a validade dos decisores na indicação das soluções finais do problema de aprendizagem sob a abordagem multiobjetivo.

Ainda neste capítulo, serão vistos os resultados da decisão de um decisor hipotético baseado na medida da distância euclidiana em relação a uma provável solução utópica, uma solução com os mínimos absolutos de cada objetivo individual: erro e norma iguais a zero. Os resultados foram inseridos nos gráficos gerados nessas simulações com o objetivo de analisar alguma relação entre as soluções contidas no conjunto Pareto-ótimo e a solução utópica, comumente caracterizada na abordagem multiobjetivo de otimização.

7.1 Análise de Ruído

Nesta sessão foi realizado um conjunto de experimentos com o treinamento de uma rede MLP para problemas de aproximação de funções utilizando diferentes estratégias de decisão no conjunto Pareto-ótimo. Os experimentos realizados utilizaram diferentes níveis de ruídos. O objetivo desses experimentos é analisar a influência do ruído presente no conjunto de dados na decisão por um modelo no conjunto de soluções. Os experimentos foram realizados com 100 pontos amostrados da função geradora e acrescidos de um ruído gaussiano de média $\mu = 0$ e variâncias $\sigma^2 = 0.1, 0.2$ e 0.3. Um conjunto de teste composto de 500 dados foi utilizado para comparar o desempenho das soluções de cada estratégia de decisão. A rede MLP utilizada em todas simulações possui a seguinte arquitetura: 1 entrada, 20 neurônios na camada escondida, 1 saída. As funções de ativação são sigmoidal (camada escondida) e linear (camada saída). O treinamento multi-objetivo gerou um conjunto de 20 soluções Pareto-ótimas com diferença entre normas das redes igual a 1.

O algoritmo de otimização utilizado para obter as soluções Pareto-ótimas foi o Levenberg-Marquardt adaptado por Costa et al. (2007) para treinamento de redes MLP com valores pré-definidos para a norma final dos pesos.

7.1.1 Problema $f_3(x)$

A seguir, os resultados das simulações com uma função acrescida de um ruído gaussiano $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x}) + \xi$

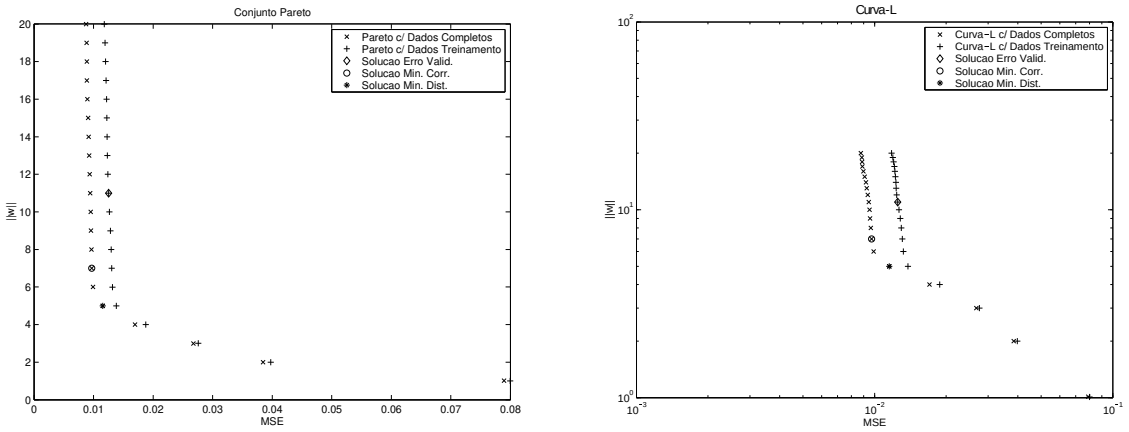


Figura 7.1: Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.1$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso.

Os experimentos realizados para este problema foram os que conseguiram distiguir melhor as decisões de cada estratégia. De forma geral, a solução obtida pelo decisor de mínimo erro de validação foi a que mais apresentou dificuldades em conseguir obter uma boa solução. Na maioria das situações simuladas (variação do ruído) o decisor por mínimo erro de validação apresentou uma solução sobre-ajustada ao problema. As demais estratégias de decisão, todas que não fazem uso de dados de validação, conseguiram encontrar a solução mais adequada ou próxima da mais adequada nas condições de teste utilizadas.

Com este experimento, percebeu-se a limitação do decisor baseado em erro de validação quando a amostra utilizada para essa tarefa era insuficiente representar a informação útil. Isto evidenciou uma das fragilidades inerentes a abordagem de decisão baseada em dados de validação.

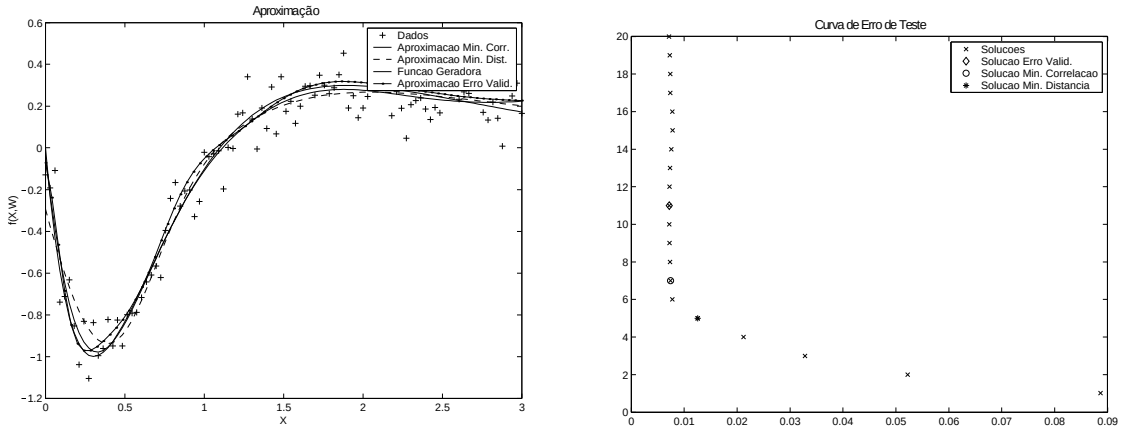


Figura 7.2: Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.1$. Este conjunto de teste é composto por 500 dados.

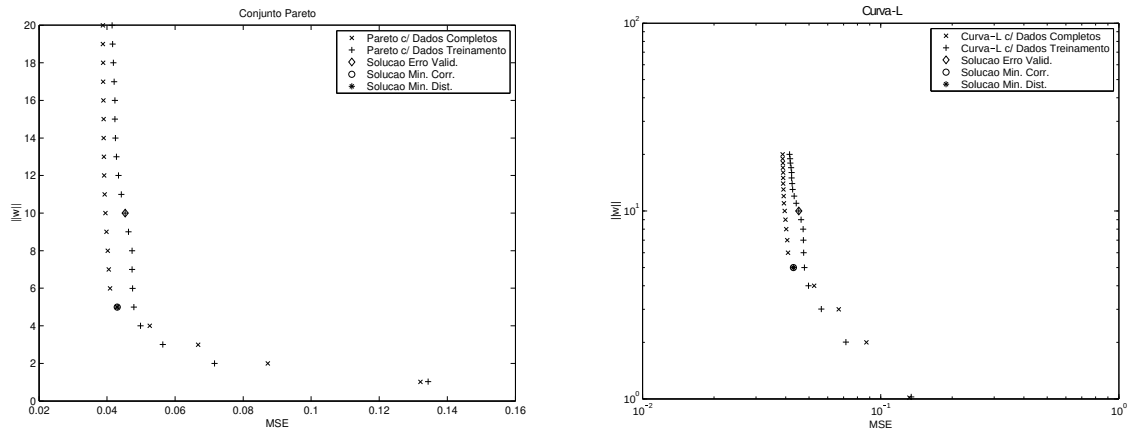


Figura 7.3: Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.2$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso.

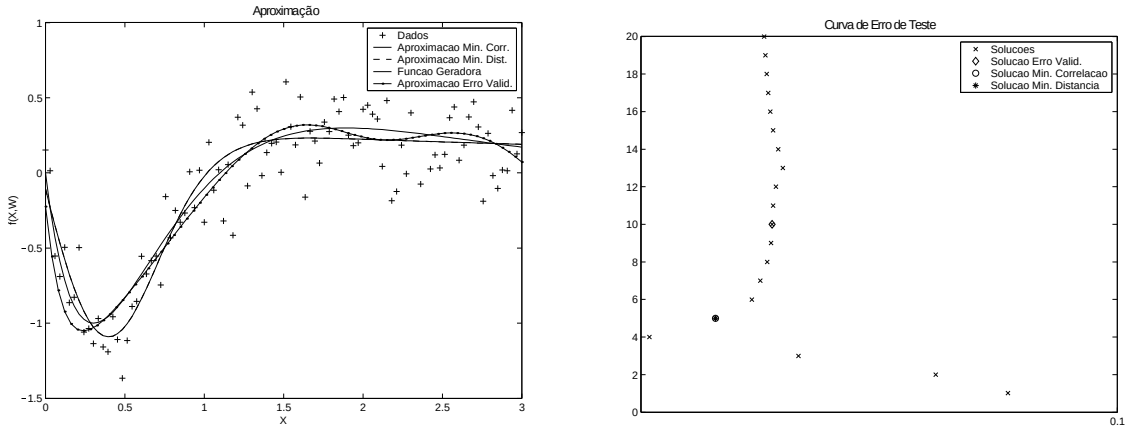


Figura 7.4: Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.2$. Este conjunto de teste é composto por 500 dados.

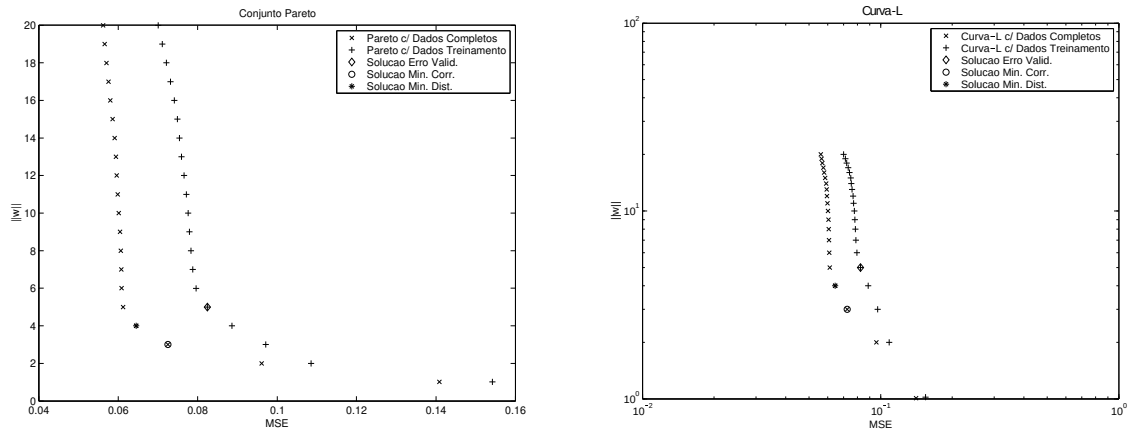


Figura 7.5: Conjunto de Pareto e Curva-L com as soluções dos decisores para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.3$. Há um conjunto Pareto-ótimo obtido com o treinamento usando parte do conjunto de treinamento (MOBJ com decisor por Validação) e um outro com o treinamento usando todos os dados (100 dados). O mesmo foi feito com a Curva-L. As soluções dos 3 decisores são indicadas em cada caso.

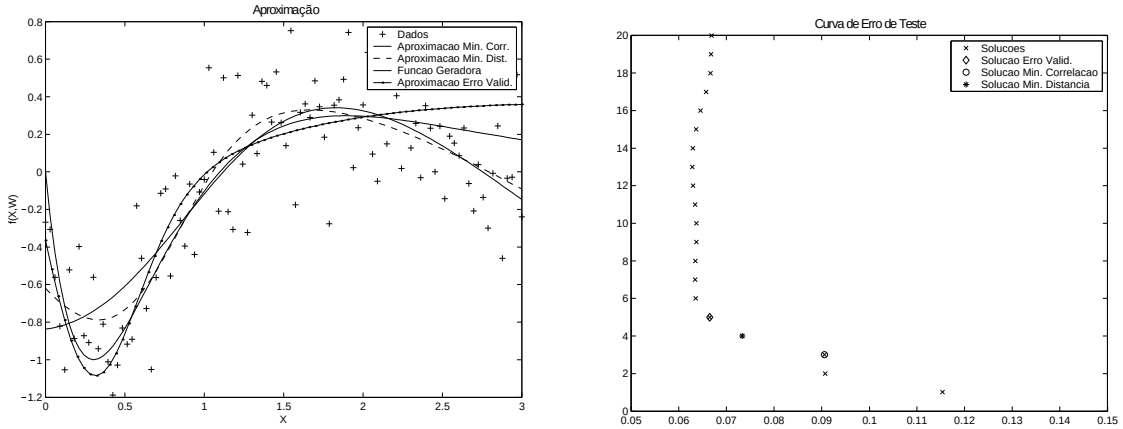


Figura 7.6: Aproximações obtidas pelas indicações dos decisores e a Curva de Teste: erro de teste x norma para o problema $f_3(x)$ com ruído de $\sigma^2 = 0.3$. Este conjunto de teste é composto por 500 dados.

7.1.2 Análise de Ruído Correlacionado

Neste momento são apresentados os resultados da estratégia de decisão por mínima correlação em problemas de regressão com um ruído correlacionado nos dados. O objetivo é destacar a eficiência das estratégias de decisão em um cenário um pouco mais complexo, onde o ruído foge das características esperadas pela estratégia de decisão por mínima autocorrelação.

A Função de Aproximação

A função $f_2(x) = \frac{(x-2)(2x+1)}{(1+x^2)} + \xi_{corr}$ corresponde a uma função com um ruído ξ_{corr} correlacionado com uma função seno gerado da seguinte forma, segundo a equação 7.1:

$$\xi_{corr}(i) = 0.9 * sin(x(i-1)) + \sigma * \xi(i-1); \tag{7.1}$$

onde x é a entrada da função $f_2(x)$, correspondente ao intervalo $[-\pi,\ \pi]$, o ruído ξ corresponde ao ruído branco gerado com média 0 e variância $\sigma^2 = 0.1$ e ξ_{corr} é o ruído correlacionado gerado a partir do seno de x e do ruído branco. A Equação 7.1 mostra como foi gerado este ruído.

Algumas figuras descrevem a distribuição dos dados usados para construção do problema de regressão com a função sinc.

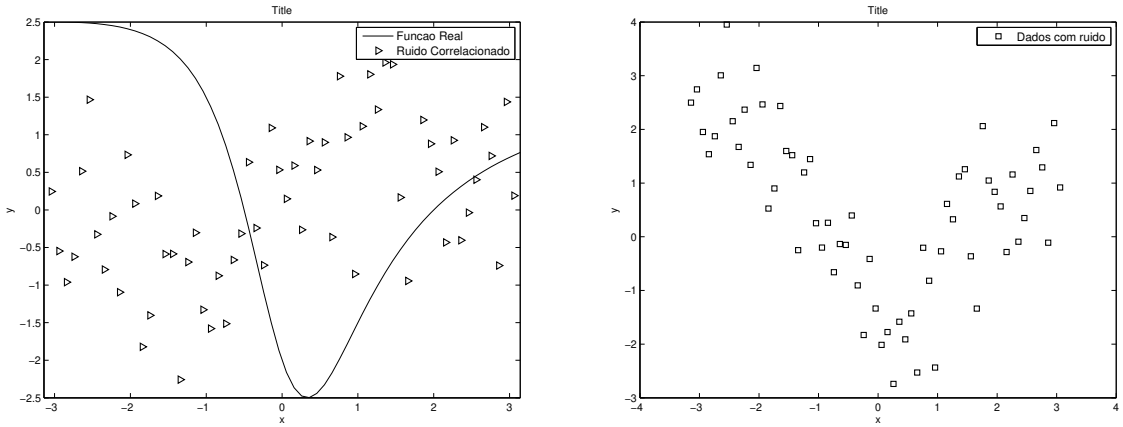


Figura 7.7: Função geradora $\frac{(x-2)(2x+1)}{(1+x^2)}$ com o ruído correlacionado ξ_{corr}, exibidos separadamente, e as amostras da função com ruído.

A Figura 7.8 e 7.9 permitem uma análise do nível de correlação, seja positiva ou negativa, entre os dados para diferentes *lags*. A função de autocorrelação do ruído não correlacionado apresenta um pico e os valores restante bem próximos de nulo. A função de autocorrelação de um ruído correlacionado apresenta um pico e os valores restantes da autocorrelação caem gradativamente à medida de distanciam-se do pico.

O treinamento multi-objetivo com decisor por mínima correlação recebeu um conjunto de dados com o ruído correlacionado inserido conforme a Equação 7.1. Como pode ser

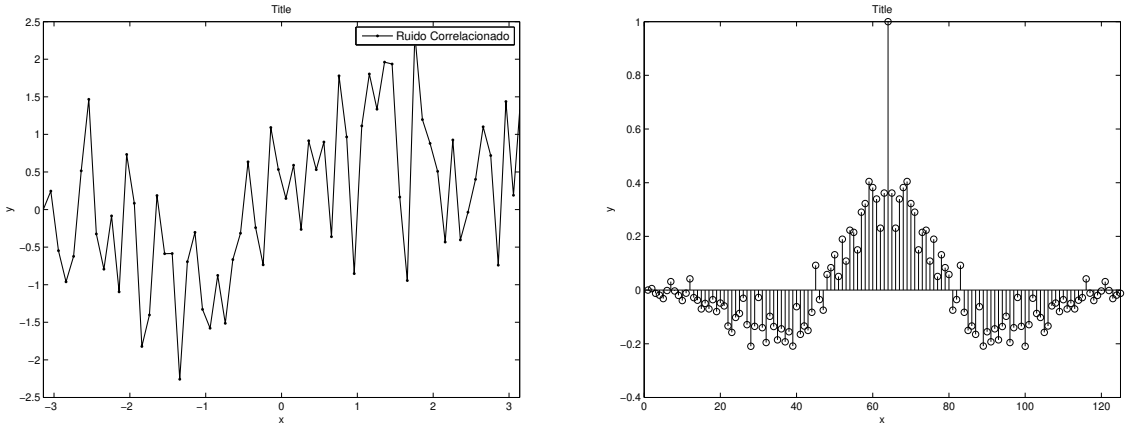


Figura 7.8: O ruído correlacionado gerado e a função de autocorrelação deste ruído.

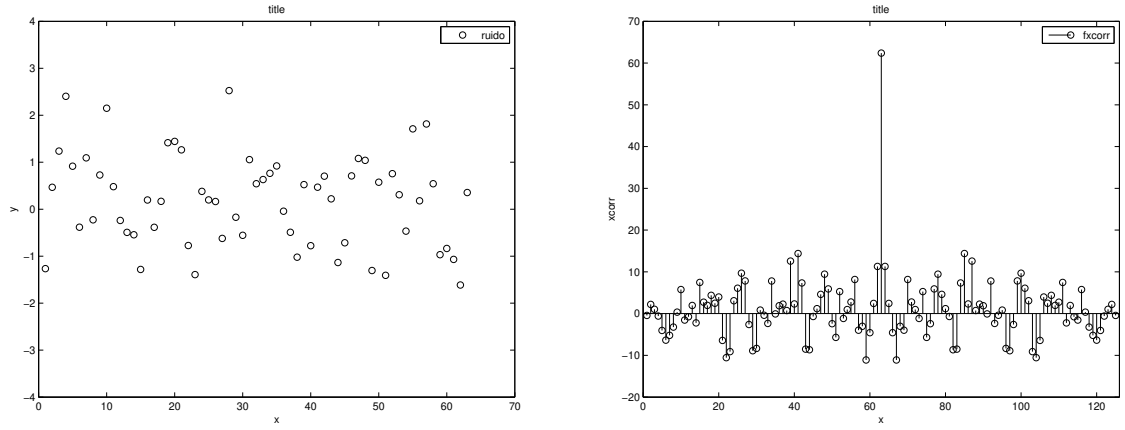


Figura 7.9: O ruído branco e a função de autocorrelação deste ruído.

visto, pela Figura 7.8 o conjunto de dados de treinamento possuía um ruído correlacionado para testar o desempenho do decisor por mínima correlação nesse cenário.

A Figura 7.10 mostra que a solução obtida pelo decisor de mínima correlação localiza-se acima da região de máxima curvatura da *Curva L* que, segundo sua teoria, indicaria a região onde está contida a solução de maior capacidade de generalização. Mesmo com um ruído com elevado índice de correlação, a tomada de decisão sofreu baixa interferência deste fator. Isto pode ser percebido pela localização da solução MOBJ de mínima correlação se comparada com a região de máxima curvatura da *Curva L*.

A Figura 7.11 mostra o erro das soluções MOBJ em relação a aproximação da função geradora dos dados, livre de ruído. O resultado apresentado na Figura 7.11 destacou a melhor solução e sua localização em relação a região de máxima curvatura da *Curva L*. Portanto neste experimento a *Curva L* mostrou-se consistente e o decisor por mínima correlação sofreu uma baixa interferência do ruído, indicando uma solução com efeito de *overfitting*. Ainda assim, respondendo de modo aceitável para problemas com ruído

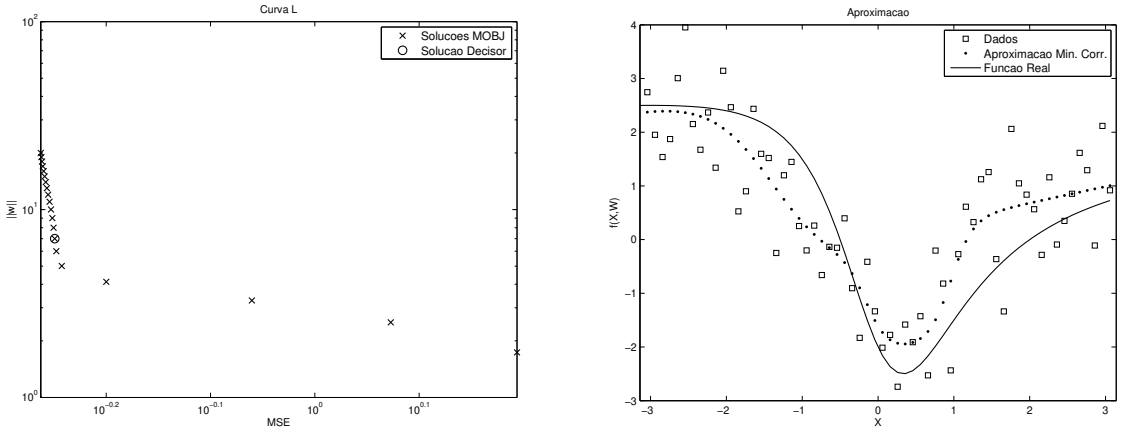


Figura 7.10: A *Curva L* com a decisão por mínima correlação e aproximação correspondente à solução escolhida por mínima correlação com dados de ruídos correlacionados.

correlacionado.

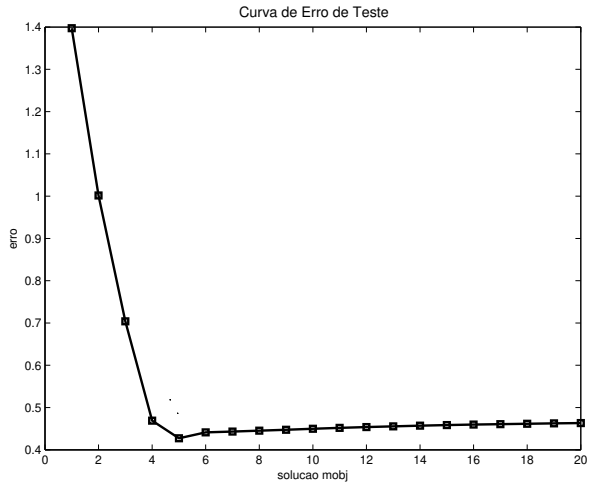


Figura 7.11: A Curva com os erros de teste de cada solução Pareto-ótima em relação à dados que não foram corrompidos com ruído. Aqui a solução com mínimo erro indica a localização da melhor aproximação da rede MLP em relação ao conjunto de dados disponível neste caso.

Aqui foram realizados experimentos com o treinamento de uma rede MLP para problemas de aproximação de funções utilizando diferentes estratégias de decisão no conjunto Pareto-ótimo. As simulações foram realizadas com conjunto de dados amostrados sob diferentes densidades. O objetivo desses experimentos era avaliar a influência do número de dados disponível na construção do conjunto de soluções Pareto-ótimas e, também da escolha da melhor solução. Nos experimentos realizados utilizou-se um conjunto de dados de 200 pontos para que o treinamento multi-objetivo e um conjunto de dados de teste com 500 pontos para avaliar o desempenho das soluções indicadas pelas estratégias de decisão. Os experimentos foram realizados adicionando um ruído gaussiano de média $\mu = 0$ e variâncias $\sigma^2 = 0.1$ ao conjunto de dados de treinamento e de teste. A rede MLP utilizada em todas simulações possui a seguinte arquitetura: 1 entrada, 20 neurônios na camada escondida, 1 saída, as funções de ativação são sigmoidal (camada escondida) e linear (camada saída). O treinamento multi-objetivo gerou um conjunto de 20 soluções Pareto-ótimas com diferença entre normas das redes igual a 1. O algoritmo de otimização utilizado para obter as soluções Pareto-ótimas é o Levenberg-Marquardt adaptado por Costa et al. (2007) para treinamento de redes MLP com valores pré-definidos de norma dos pesos.

7.1.3 Problema $f_1(x) = sen(x) + \xi$

A seguir, os resultados das simulações com as estratégias de decisão com três conjuntos de dados disponível, o primeiro com somente 15 dados para treinamento (e 300 dados para teste), um com 100 dados para treinamento (e 200 dados para teste) e outro com 200 dados para treinamento(e 500 dados para teste). As figuras 7.12 à 7.13 mostram os resultados do treinamento multi-objetivo com 15 dados. As figuras 7.14 à 7.15 mostram os resultados do treinamento multi-objetivo com 100 dados. As figuras 7.16 à 7.17 mostram os resultados do treinamento multi-objetivo com 200 dados.

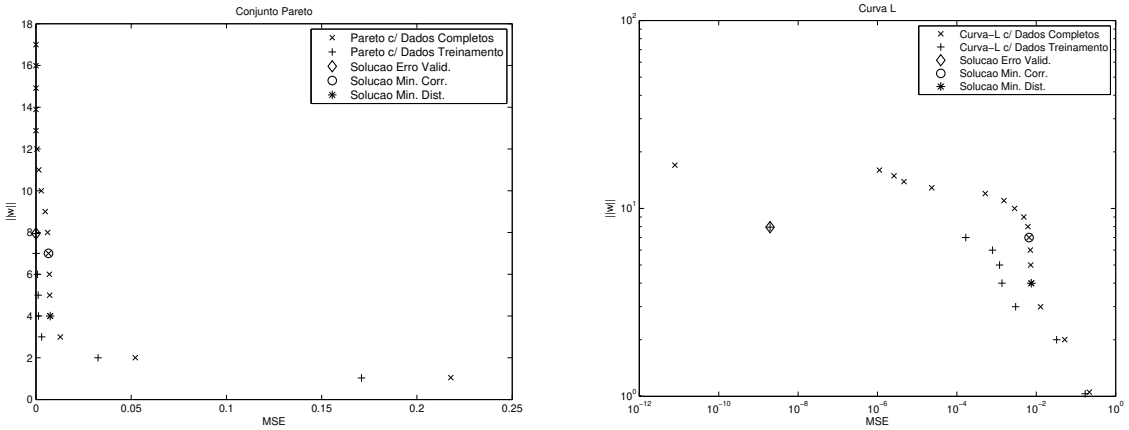


Figura 7.12: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 15 dados de treinamento.

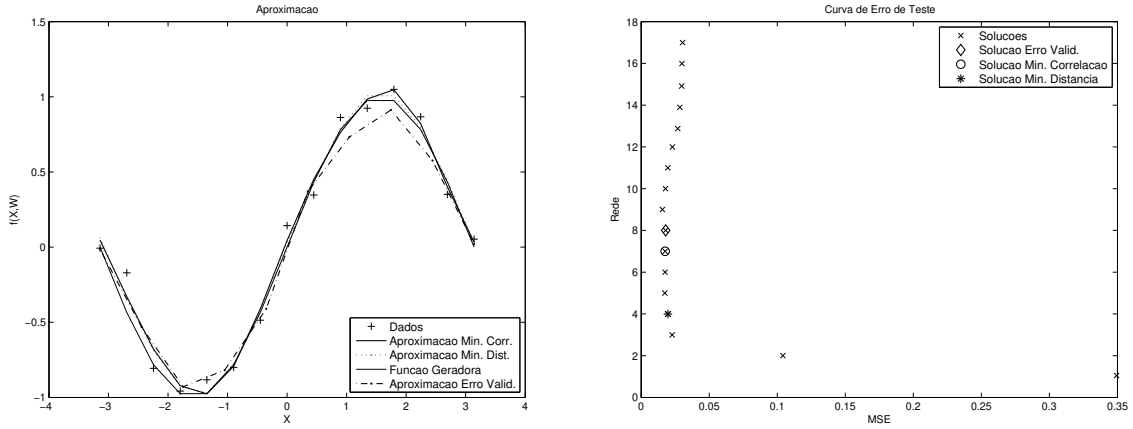


Figura 7.13: Aproximações obtidas com o treinamento usando 15 dados e a curva do erro de teste x norma utilizando 300 dados de teste.

Os resultados apresentados nas Figuras 7.12 e 7.13 destacam a capacidade do decisor por mínima correlação em trabalhar com problemas de aproximação com amostras muito reduzidas. As Curvas Pareto-ótima e a *Curva L* mostram que o treinamento multiobjetivo é favorecido quando todas as amostras disponíveis podem ser usadas para o treinamento, o que ocorre com a tomada de decisão baseada na autocorrelação do resíduo. A curva com os erros de teste, da Figura 7.13, destacou a localização da região de soluções MOBJ com melhor aproximação dos dados de teste.

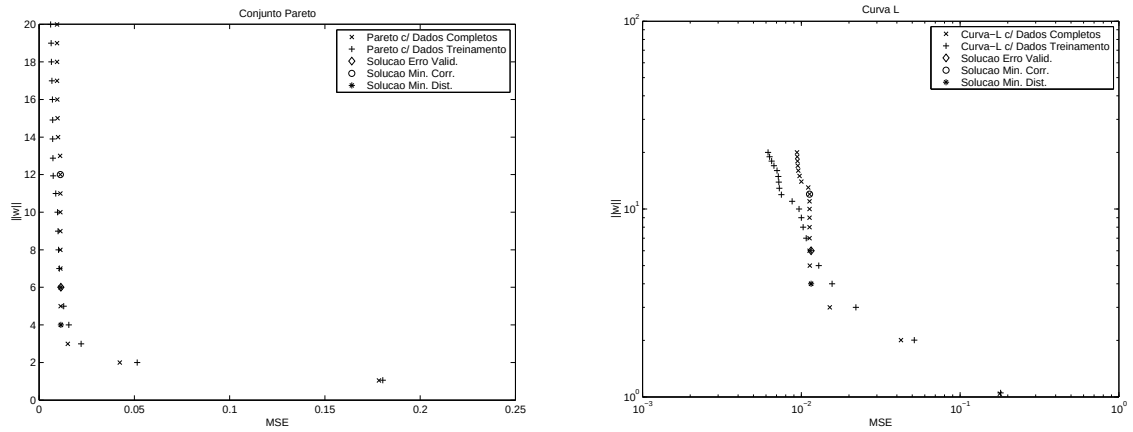


Figura 7.14: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento.

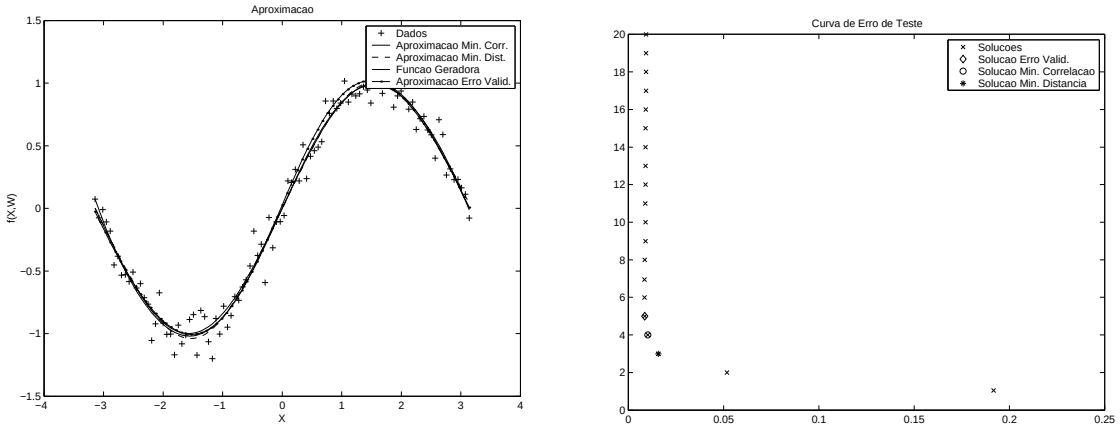


Figura 7.15: Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste.

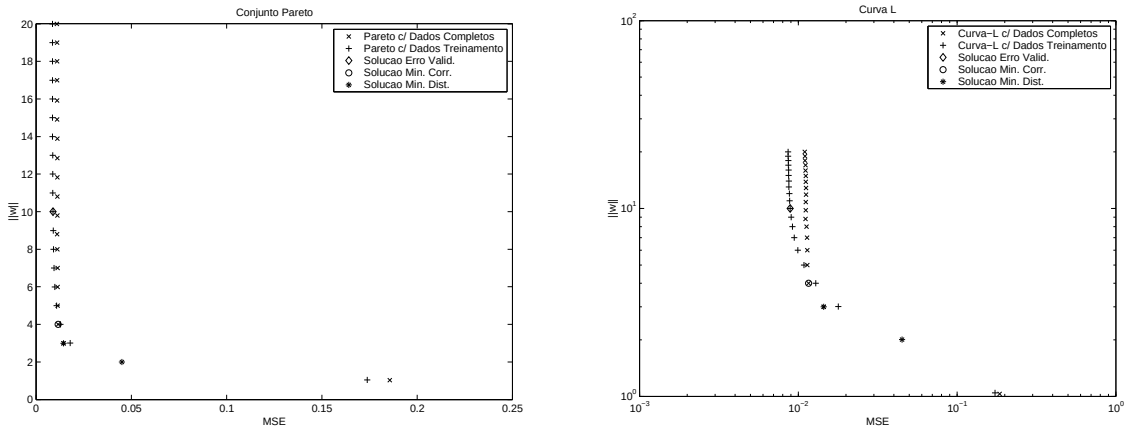


Figura 7.16: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento.

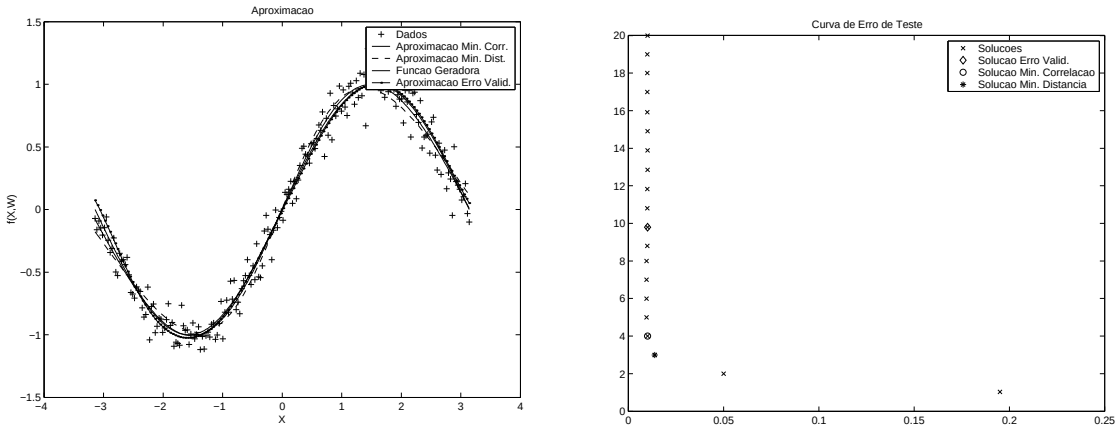


Figura 7.17: Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste.

7.1.4 Problema $f_2(x) = \frac{(x-2)(2x+1)}{(1+x^2)} + \xi$

Aqui os resultados das simulações com as estratégias de decisão com dois conjuntos de dados, um com 100 dados para treinamento e 200 dados para teste enquanto outro com 200 dados para treinamento e 500 dados para teste. As figuras 7.20 à 7.21 mostram os resultados do treinamento multi-objetivo com 100 dados. As figuras 7.22 à 7.23 mostram os resultados do treinamento multi-objetivo com 200 dados.

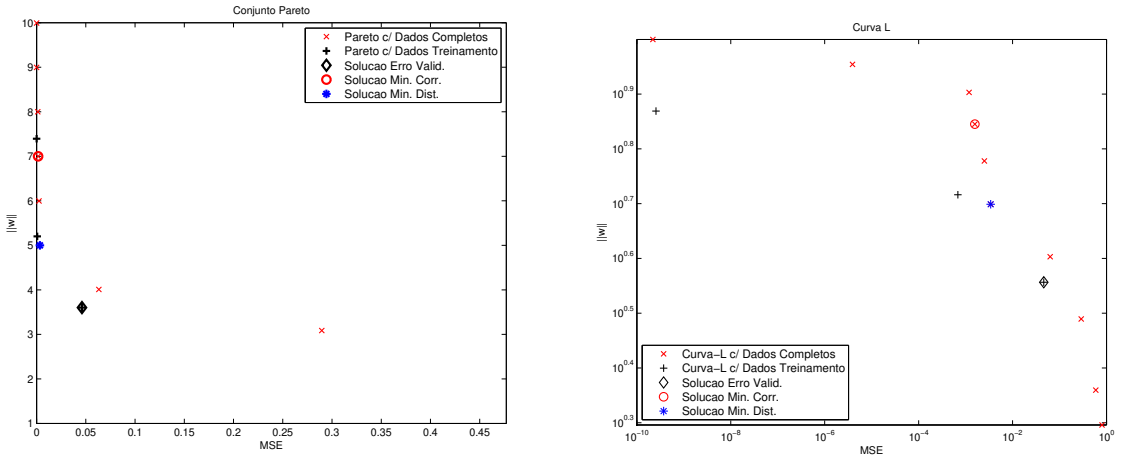


Figura 7.18: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 15 dados de treinamento.

Os resultados apresentados nas Figuras 7.18 e 7.19 destacaram a precisão do decisor por mínima correlação em trabalhar com problemas de aproximação amostras com muito reduzidas. Nesta simulação, a decisão por erro de validação, afetada pela reduzida quantidade de amostras disponíveis, selecionou uma rede MOBJ com norma muito baixa, conforme mostra a Figura 7.18. As Curva de Pareto e *Curva L* mostram que o treinamento multi-objetivo foi favorecido quando todas as amostras disponíveis puderam ser usadas para o treinamento, o que ocorreu com a tomada de decisão baseada na autocorrelação do resíduo. A Curva de erros de teste, da Figura 7.19, destaca a localização da região de soluções MOBJ com melhor aproximação dos dados de teste.

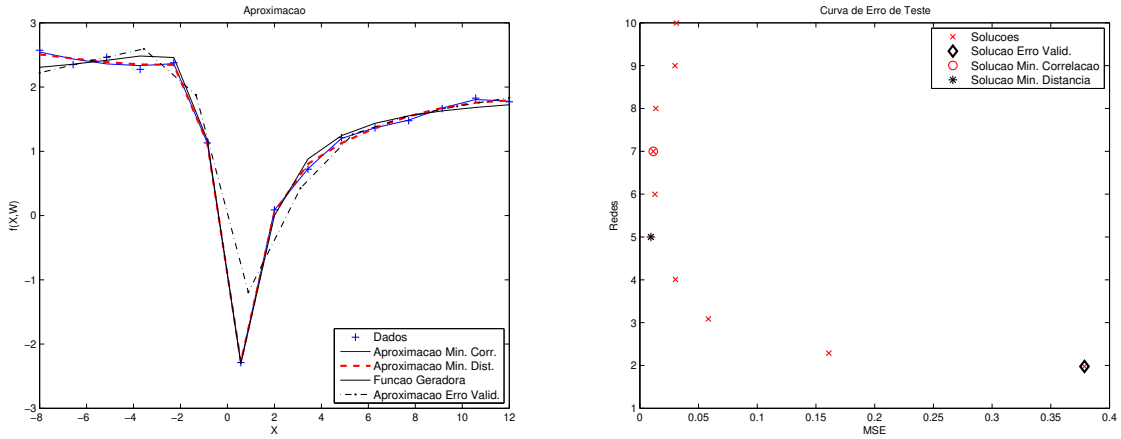


Figura 7.19: Aproximações obtidas com o treinamento usando 15 dados e a curva do erro de teste x norma utilizando 200 dados de teste.

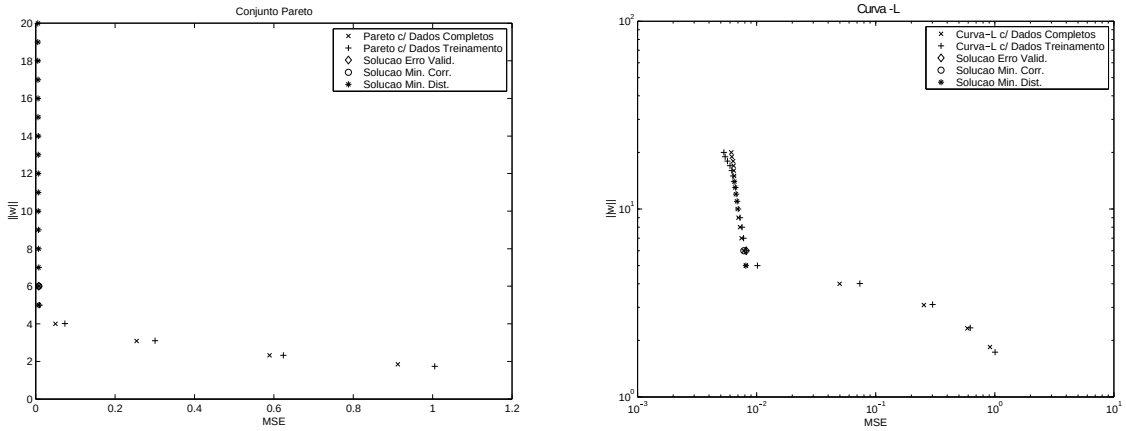


Figura 7.20: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento.

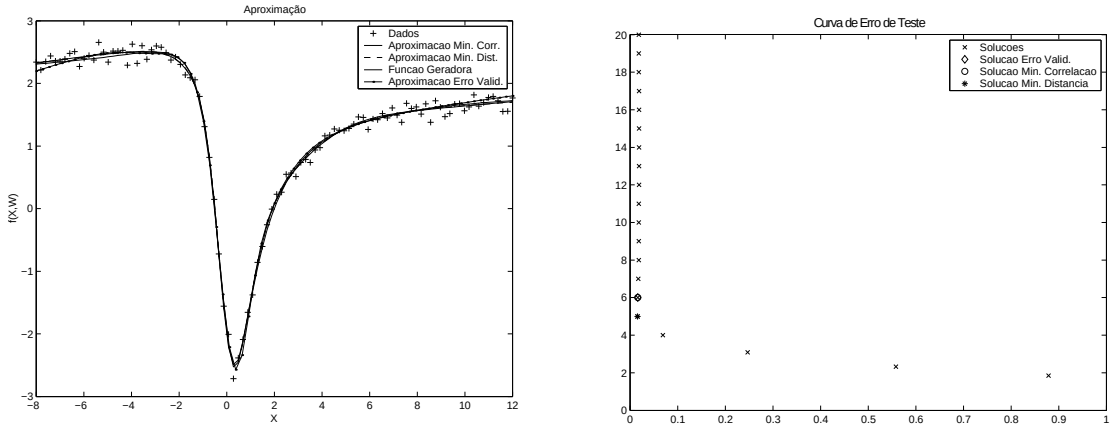


Figura 7.21: Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste.

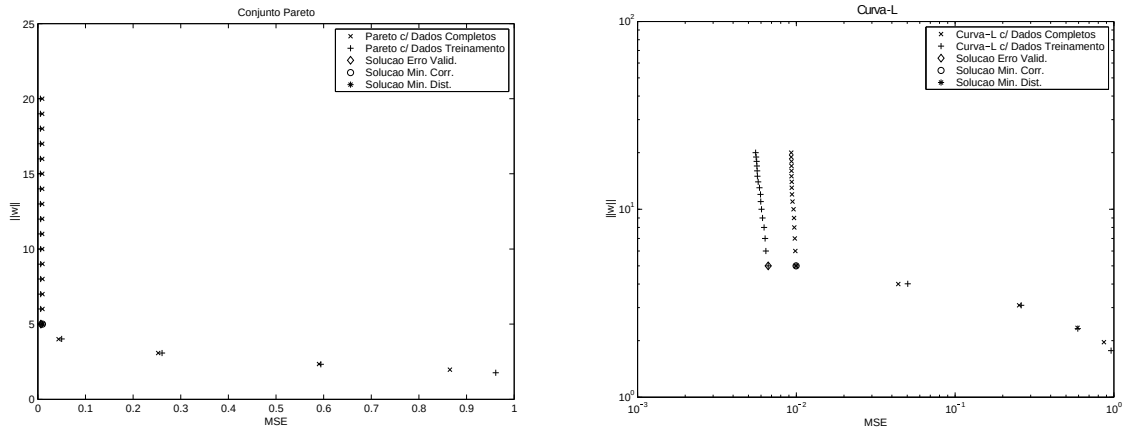


Figura 7.22: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento.

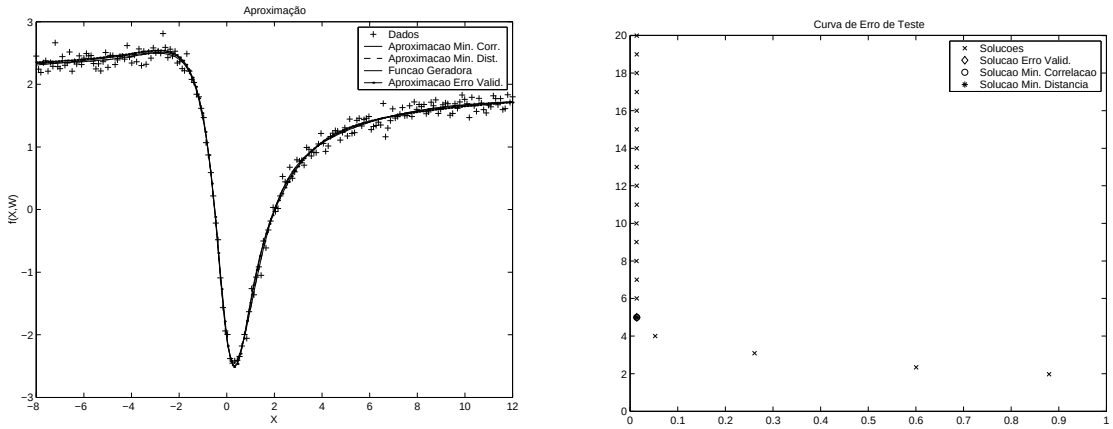


Figura 7.23: Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste.

7.1.5 Problema $f_3(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x}) + \xi$

A seguir, os resultados das simulações com as estratégias de decisão com dois conjuntos de dados, um com 100 dados para treinamento e 200 dados para teste enquanto outro com 200 dados para treinamento e 500 dados para teste. As figuras 7.24 à 7.25 mostram os resultados do treinamento multi-objetivo com 100 dados. As figuras 7.26 à 7.27 mostram os resultados do treinamento multi-objetivo com 200 dados.

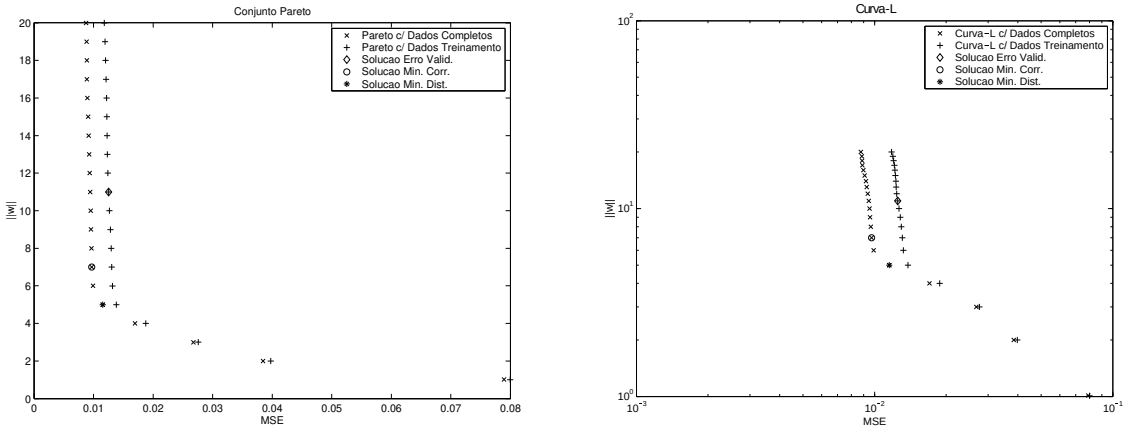


Figura 7.24: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 100 dados de treinamento.

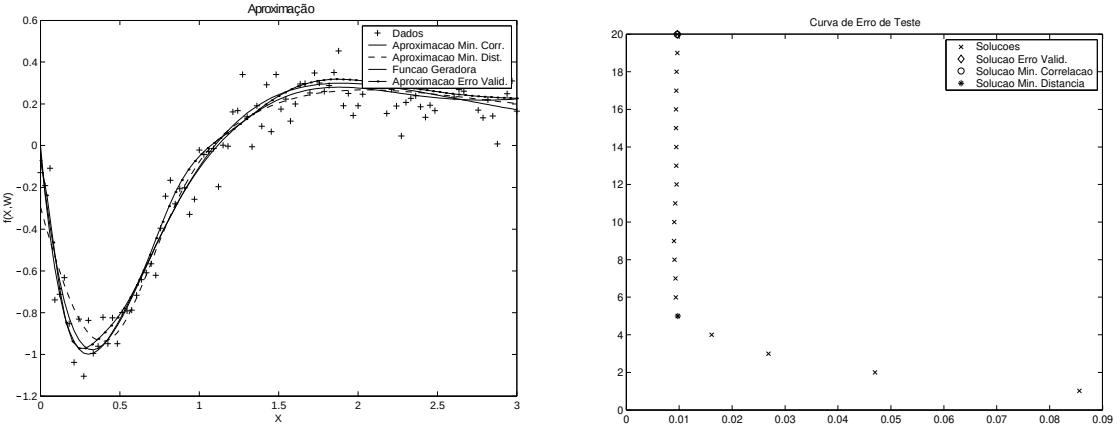


Figura 7.25: Aproximações obtidas com o treinamento usando 100 dados e a curva do erro de teste x norma utilizando 200 dados de teste.

O conjunto de experimentos no qual foi variado o número de dados em cada problema mostrou que, de forma geral, não houve grande mudança no comportamento dos decisores. A oscilação da solução indicada em cada decisor com o aumento do número de pontos não causou grande mudança na qualidade da resposta obtida pelos modelos. O teste com

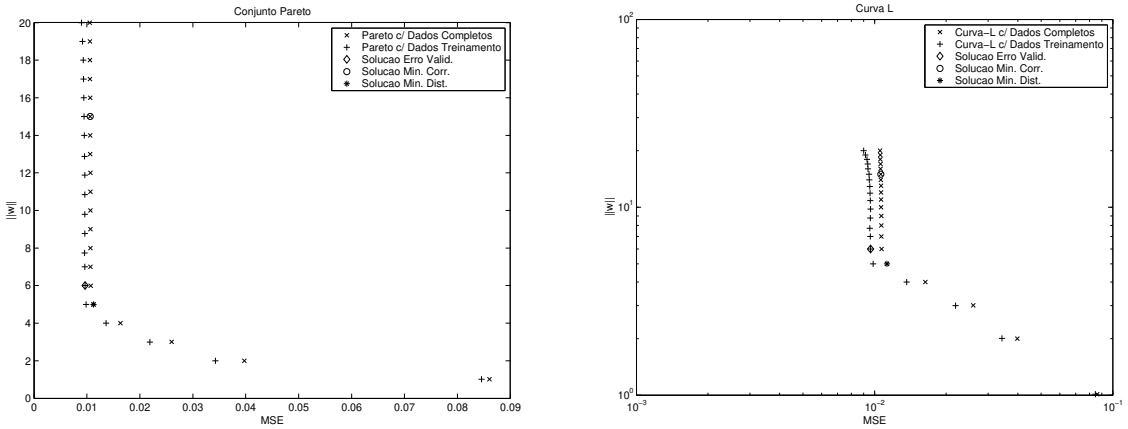


Figura 7.26: Conjunto de Pareto e Curva-L com as soluções dos decisores utilizando uma amostra com 200 dados de treinamento.

a variação da intensidade do ruído foi bem mais discriminante. Mas é válido analisar ainda as condições de sub-amostragens para os problemas de modo a analisar o limite inferior no qual cada decisor ainda é capaz de obter a solução válida.

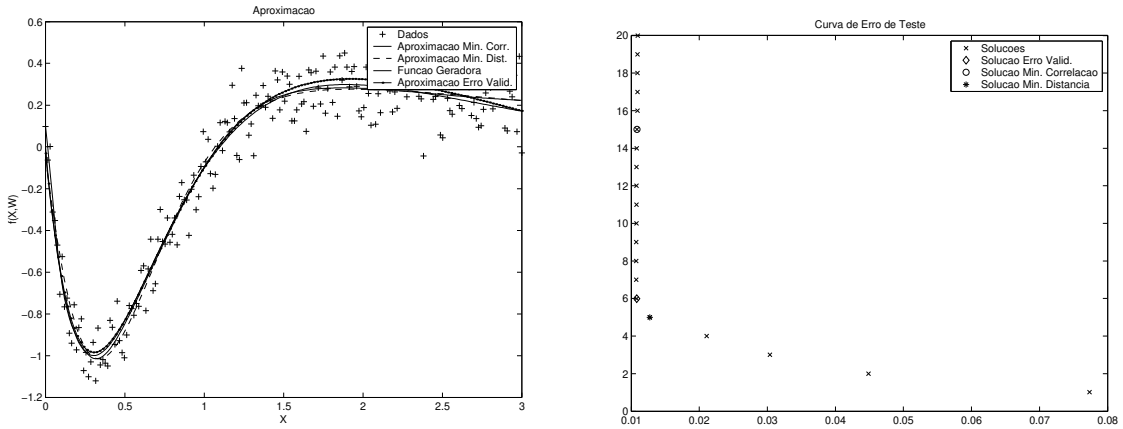


Figura 7.27: Aproximações obtidas com o treinamento usando 200 dados e a curva do erro de teste x norma utilizando 500 dados de teste.

7.2 Comparação entre decisores

Nesta seção serão apresentados os resultados de experimentos que mediram a precisão da solução escolhida num conjunto de dados de teste com muitos pontos amostrados. Com isso é possível analisar numericamente a qualidade de cada modelo escolhido.

Nos experimentos realizados foi utilizado um conjunto de 200 dados para o treinamento e um conjunto de teste com 2000 dados para avaliar o desempenho das soluções de cada estratégia de decisão por meio do erro médio quadrático (MSE) das soluções em relação ao conjunto de teste. Os experimentos foram realizados com um ruído gaussiano de média $\mu = 0$ e variâncias $\sigma^2 = 0.1$ inserido nos dados. A rede MLP utilizada em todas simulações possuía a seguinte arquitetura: 1 entrada, 20 neurônios na camada escondida, 1 saída, as funções de ativação são sigmoidal (camada escondida) e linear (camada saída). O treinamento multi-objetivo gerou um conjunto de 20 soluções Pareto-ótimas com diferença entre normas das redes igual a 1. O algoritmo de otimização utilizado para obter as soluções Pareto-ótimas é o Levenberg-Marquardt adaptado em Costa et al. (2007) para treinamento de redes MLP com valores pré-definidos de norma dos pesos.

As Tabelas 7.1 e 7.2 apresentam os erros de cada modelo escolhido para o conjunto de teste nos dois problemas teste simulados.

Tabela 7.1: Erro (MSE) no conjunto de teste de 2000 dados - Problema $f_2(x)$

Decisor	erro (MSE)
Validação	0.0077
Correlação	0.0077
Curva - L	0.0077

Tabela 7.2: Erro (MSE) no conjunto de teste de 2000 dados - Problema $f_3(x)$

Decisor	erro (MSE)
Validação	0.0123
Correlação	0.0133
Curva - L	0.0151

Analisando os resultados das Tabelas 7.1 e 7.2 as estratégias de decisão adotadas nas simulações não apresentaram diferenças significativas no desempenho de aproximação da função geradora de cada problema. Os resultados tendem a garantir que a tendência das estratégias é convergir para uma mesma região de soluções para grandes conjuntos de dados de treinamento. Os resultados obtidos com a decisão por meio do critério da Curva-l apesar de apresentarem bons resultados nesses experimentos, podem não gerar soluções satisfatórias quando a curva obtida não apresentar o formato aproximado da letra L, tornando mais difícil o processo de identificação do *corner*, já que em alguns

problemas, este *corner* pode não ser evidenciado.

Os resultados aqui mostraram uma equivalência entre as respostas das estratégias baseadas em erro de validação e mínima autocorrelação. Experimentos utilizando conjuntos de dados reduzidos tornariam menos eficiente a qualidade do modelo obtido através da decisão por mínimo erro de validação. Isto era esperado uma vez que utilizando essa estratégia, é necessário separar uma fração dos dados disponíveis para realizar o processo de decisão. Diminuindo assim o volume de dados amostrados para o aprendizado. Os conjuntos Pareto-ótimo gerados com o treinamento usando uma amostragem reduzida apresentam maior deslocamento no sentindo em que os erros médios quadráticos das soluções aumentam, como pôde ser visto nos gráficos das curvas de Pareto desse capítulo.

7.3 Comentários Finais

Os resultados obtidos neste capítulo validam a eficácia das estratégias de decisão em problemas de regressão segundo a abordagem multiobjetivo. Foi possível destacar que a estratégia de decisão por mínima autocorrelação permitia uma abordagem consistente em diferentes cenários de simulação. O princípio dessa estratégia permitiu a análise da regressão sob o ponto de vista do ruído gaussiano presente nos dados. Além disso, tal como nos experimentos em problemas de classificação, o método da *Curva L* foi usado como um indicador da região provável de serem localizadas as soluções. A caracterização da solução por mínima distância euclidiana em relação à solução utópica não mostrou-se consistente nos experimentos realizados de modo a demandar alguma formulação que comprovasse sua eficiência.

Capítulo 8

Conclusões e Propostas de Continuidade

Neste trabalho procurou-se estudar, formalizar e implementar metodologias alternativas para se realizar a tomada de decisão num conjunto reduzido de soluções, o conjunto Pareto-ótimo, em problemas de aprendizagem de máquina multiobjetivo. O objetivo do trabalho era estudar, propor e elaborar estratégias capazes de encontrar soluções que melhor representassem o equilíbrio entre os efeitos de variância e polarização em modelos de máquinas de aprendizagem. As estratégias projetadas destacaram-se por basear-se em princípios que as tornavam independentes de amostras. Para isso, utilizaram-se de informações sobre a estrutura do tipo de problema de aprendizagem para orientar o método de decisão.

Ambas as estratégias desenvolvidas mostraram-se eficientes para as classes de problemas de aprendizado para as quais foram testadas. A estratégia de decisão baseada na medida da mínima auto-correlação residual garantia que o modelo escolhido representava a melhor aproximação da função real por meio da suposição de dados com ruído e a natureza gaussiana da distribuição desses ruídos. O princípio do decisor de mínima auto-correlação do ruído mostrou-se eficiente mesmo quando o ruído poderia apresentar-se correlacionado com os dados. Além disso, essa estratégia de decisão também mostrou-se eficiente quando o número de exemplos de treinamento era reduzido. Essa estratégia limitou-se aos problemas de regressão (ou aproximação de funções), uma vez que os dados em um problema deste tipo apresentavam uma auto-correlação temporal e somente nessas condições era possível realizar a análise do resíduo e sua auto-correlação.

A estratégia de decisão por teste de hipótese garantiu que a solução escolhida reproduzisse a incerteza nos rótulos (ou classes) dos dados, evitando o efeito de variância causado pelo ajuste excessivo no treinamento e por uma eventual escolha polarizada ao usar o decisor com dados de validação. Os resultados apresentados por este decisor destacaram a influência que o conhecimento prévio de um especialista do domínio pode ter na tomada de decisão. Essa estratégia mostrou-se importante em situações que há algum conhecimento sobre o processo de modo que pudesse ser adaptado ao método MOBJ. Essa abordagem também mostrou-se consistente diante de diferentes circunstâncias apresen-

tadas nas simulações.

Neste trabalho procurou-se evidenciar que as novas estratégias desenvolvidas não contrariavam os princípios dos decisores anteriormente desenvolvidos, mas ampliavam o conjunto de técnicas desenvolvidas para o método multiobjetivo de aprendizagem de máquina. Objetivou-se destacar que o processo de tomada de decisão original necessitava de uma amostragem do conjunto de dados disponível para o processo de aprendizagem de modo a inferir a solução do conjunto Pareto-ótimo que lhe parecesse a mais propícia. Então o valor de se desenvolver novas abordagens de decisão independentes de amostragem está na consistência da lógica do processo decisório, que guiará o algoritmo de modo sistemático e bem definido a indicar sua solução ou região de soluções sem estar sob a influência da qualidade dos dados que até então eram necessários.

O uso da *Curva L* no método MOBJ permitiu que fosse avaliada como essa abordagem poderia indicar a região de soluções onde haveria um equilíbrio entre o erro de ajuste aos dados e a complexidade do modelo. Sua formulação originalmente direcionada para modelos lineares de ajuste aos dados permitiu um olhar particular em relação abordagem multiobjetivo que poderia ter o conjunto Pareto-ótimo representado graficamente tal como a *Curva L* e, por ela identificar uma provável solução ou soluções equivalentes. Com sua fundamentação teórica desenvolvida para os métodos de regularização a *Curva L* já era um método bem conhecido e de bons resultados quando utilizada em em modelos lineares, cujo objetivo era identificar a intensidade do parâmetro regularização.

No entanto, na tentativa de avaliar esta curva sob a perspectiva do método MOBJ, suas características de indicação de uma região de máxima curvatura onde localizaria-se a melhor solução para o problema acabou não sendo consistente para Redes MLP e a abordagem de construção do conjunto Pareto-ótimo via programação por metas. Em razão das características de não convexidade da curva Pareto-ótima obtida a característica forma da *Curva L* nem sempre era evidente. Isso permite que todos os métodos desenvolvidos aqui também possam ser utilizados nos métodos de regularização com boa expectativa de estimativa do parâmetro de regularização.

Propostas de Continuidade

Sugere-se como propostas para continuação deste trabalho de tese, investir nos seguintes problemas relacionados ao tema:

- Caracterizar melhor o critério de medida da complexidade de modelos de aprendizado por meio de seu comportamento. Determinação das componentes de alta frequência do sinal de resposta do modelo de aproximação.

- Implementar novas estratégias de decisão que não utilizem informação de dados de validação. Isto implica na proposta de estudos da Regularização Bayesiana e seu critério de decisão bayesiano da importância de cada objetivo durante o processo de aprendizado. Projetar um método MOBJ com critério bayesiano para direcionar a determinação da(s) solução(ões) Pareto-ótima(s).

- Conhecida as limitações de métodos de regularização (ex.: *Curva-L*, Regularização Bayesiana, *Weight Decay*) em mapear regiões não-convexas da fronteira Pareto-ótima. É importante compreender, propor e implementar estratégias de modificação da função de custo ponderada desses métodos de modo que seja possível atingir pontos, no espaço dos objetivos, que sejam não-convexos. Alguns métodos para modificação da função de custo ponderada já foram desenvolvidos na literatura e os mesmos eram capazes de mapear regiões não-convexas (Messac et al., 2000). O interesse nessa abordagem é que o aprendizado bayesiano dos parâmetros de regularização determine, em um único treinamento (uma única solução Pareto-ótima), a solução ótima do problema de aproximação sem precisar encontrar vários modelos de diferentes complexidades.

Referências Bibliográficas

Abbass, H. A. (2003a). Pareto neuro-evolution: Constructing ensemble of neural networks using multiobjective optimization. In Sarker, R., Reynolds, R., Abbass, H., Tan, K. C., McKay, B., Essam, D., and Gedeon, T., editors, *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003*, pages 2074–2080, Canberra. IEEE Press. 24

Abbass, H. A. (2003b). Speeding up backpropagation using multiobjective evolutionary algorithms. *Neural Computation*, 15(11):2705–2726. 23

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 14

Andonie, R. (2010). Extreme data mining: Inference from small datasets. V:280–291. 6

Asuncion, A. and Newman, D. (2007). UCI machine learning repository. 58, 80

Barroso, M. F., Takahashi, R. H., and Aguirre, L. A. (2007). Multi-objective parameter estimation via minimal correlation criterion. *Journal of Process Control*, 17(4):321–332. 28

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ. 11

Box, G. E. P. and Jenkins, G. (1990). *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated. 91

Braga, A. P., Carvalho, A. C. P. L. F., and Ludermir, T. B. (2000). *Redes Neurais Artificiais: Teoria e aplicações*. LTC (Livros Técnicos e Científicos). 262 páginas. 9, 17

Braga, A. P., Takahashi, R. H., Costa, M. A., and de Albuquerque Teixeira, R. (2006). Multi-objective algorithms for neural networks learning. In *Multi-objective machine learning*, pages 151–171. Springer. 12, 24

Burden, F. and Winkler, D. (2009). *Bayesian Regularization of Neural Networks*, pages 23–42. Humana Press, Totowa, NJ. 12

Caminhas, W. M., Vieira, D. A., and Vasconcelos, J. A. (2003). Parallel layer perceptron. *Neurocomputing*, 55(3-4):771–778. 22

Carrano, E. G., Takahashi, R. H. C., Caminhas, W. M., and Neto, O. M. (2008). A genetic algorithm for multiobjective training of anfis fuzzy networks. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 3259–3265. 5

Chen, H. and Yao, X. (2010). Multiobjective neural network ensembles based on regularized negative correlation learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(12):1738–1751. 24

Coello, C. C., Lamont, G., and van Veldhuizen, D. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Genetic and Evolutionary Computation. Springer, Berlin, Heidelberg, 2nd edition. 24

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. 12

Costa, M., Rodrigues, T., Horta, E., Braga, A., Pataro, C., Natowicz, R., Incitti, R., Rouzier, R., and Cela, A. (2009). New multi-objective algorithms for neural network training applied to genomic classification data. In *Foundations of Computational, Intelligence Volume 1*, pages 63–82. Springer. 24

Costa, M. A. and Braga, A. P. (2006). Optimization of neural networks with multiobjective lasso algorithm. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 3312–3318. IEEE. 24

Costa, M. A. and Braga, A. P. (2011). Gradient descent decomposition for multi-objective learning. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 377–384. Springer. 25

Costa, M. A., Braga, A. P., Menezes, B. R., Teixeira, R. A., and Parma, G. G. (2003). Training neural networks with a multi-objective sliding mode control algorithm. *Neurocomputing*, 51:467 – 473. 24

Costa, M. A., de Pádua Braga, A., and de Menezes, B. R. (2007). Improving generalization of mlps with sliding mode control and the levenberg-marquardt algorithm. *Neurocomputing*, 70(7):1342 – 1347. Advances in Computational Intelligence and Learning. 23, 24, 59, 95, 104, 115

Datta, S. and Das, S. (2018). Multiobjective support vector machines: Handling class imbalance with pareto optimality. *IEEE transactions on neural networks and learning systems*, 30(5):1602–1608. 22

de Albuquerque Teixeira, R., de Pádua Braga, A., Takahashi, R. H. C., and Saldanha, R. R. (2001). Recent advances in the MOBJ algorithm for training artifical neural networks. *Int. J. Neural Syst*, 11(3):265–270. 22

Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons. 22

Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34. 12, 15

Drugan, M., Wiering, M., Vamplew, P., and Chetty, M. (2017). Special issue on multi-objective reinforcement learning. *Neurocomputing*, 263:1 – 2. Multiobjective Reinforcement Learning: Theory and Applications. 23

Du, W., Leung, S. Y. S., and Kwong, C. K. (2014). Time series forecasting by neural networks: A knee point-based multiobjective evolutionary algorithm approach. *Expert Systems with Applications*, 41(18):8049–8061. 23

Edelman, A. (1988). Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560. 35

Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap. 16

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, The Netherlands. 30

Eriksson, J. and Wedin, P. (1997). Regularization methods for nonlinear least squares. 48

Ferreira, P. A. V. (1999). Otimização multiobjetivo: Teoria e aplicações. Tese de Livre Docência. 19

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and bias/variance dilemma. *Neural Computing*, 4(1):1–58. 3, 4

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269. 12

Gold, C. and Sollich, P. (2003). Model selection for support vector machine classification. *Neurocomputing*, 55(1-2):221–249. 13

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223. 35

Gong, Z., Chen, H., Yuan, B., and Yao, X. (2019). Multiobjective learning in the model space for time series classification. *IEEE Transactions on Cybernetics*, 49(3):918–932. 23

Graning, L., Jin, Y., and Sendhoff, B. (2006). Generalization improvement in multi-objective learning. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 4839–4846. IEEE. 23

Gu, Y., Wei, H.-L., and Balikhin, M. M. (2018). Nonlinear predictive model selection and model averaging using information criteria. *Systems Science & Control Engineering*, 6(1):319–328. 15

Gulliksson, M. and Wedin, P. (1998). Algorithms for using the nonlinear lcurve. 48

Handl, J. and Knowles, J. Evolutionary multiobjective clustering. In *Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature*. Springer-Verlag. 23

Handl, J. and Knowles, J. (2004). Multiobjective clustering with automatic determination of the number of clusters. Technical Report TR-COMPSYSBIO-2004-02, UMIST, Department of Chemistry. 23

Handl, J. and Knowles, J. (2006). Multi-objective clustering and cluster validation. In *Multi-Objective Machine Learning*, pages 21–47. Springer. 23

Hanke, M. (1996). Limitations of the L-curve method in ill-posed problems. *BIT*, 36:287–301. 28, 36

Hansen, P. C. (1990). Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. Sci. Statist. Comput.*, 11:503–518. 7, 35

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34:561–580. 5, 28, 35

Hansen, P. C. (2001). The L-curve and its use in the numerical treatment of inverse problems. In *Computational Inverse Problems in Electrocardiology*, number 5 in Advances in Computational Bioengineering, pages 119–142. WIT Press, Southampton. 5, 28, 35

Hansen, P. C., Jensen, T. K., and Rodriguez, G. (2004). An adaptive pruning algorithm for the discrete L-curve criterion. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby. This work was supported in part by grant no. 21-03-0574 from the Danish Natural Science Research Foundation, and by COFIN grant no. 2002014121 from MIUR (Italy). 35

Hansen, P. C. and O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14:1487–1503. 5, 36, 37

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA. 4, 11, 13

Hofmann, B. (1995). Ill-posedness and regularization of inverse problems–A review of mathematical methods. In Lübbig, H., editor, *The Inverse Problem*, pages 45–66. Akademie Verlag, Berlin. 5

Ishibuchi, H. and Nojima, Y. (2007). Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, 44(1):4–31. 24

Jin, Y., editor (2006). *Multi-objetive Machine Learning*. Springer, 1 edition. 24

Jin, Y. and Sendhoff, B. (2008). Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):397–415. 24

Kaoutar, S. and Mohamed, E. (2017). Multi-criteria optimization of neural networks using multi-objective genetic algorithm. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–4. 23

Kindermann, S. and Raik, K. (2019). A simplified l-curve method as error estimator. 28

Kokshenev, I. and Braga, A. P. (2008a). A multi-objective approach to rbf network learning. *Neurocomputing*, 71(7):1203 – 1209. Progress in Modeling, Theory, and Application of Computational Intelligenc. 5

Kokshenev, I. and Braga, A. P. (2008b). A multi-objective approach to rbf network learning. *Neurocomputing*, 71(7):1203 – 1209. Progress in Modeling, Theory, and Application of Computational Intelligenc. 22, 23

Kokshenev, I. and Braga, A. P. (2010). An efficient multi-objective learning algorithm for rbf neural network. *Neurocomputing*, 73(16):2799 – 2808. 10th Brazilian Symposium on Neural Networks (SBRN2008). 17, 22, 28

Kottathra, K. and Attikiouzel, Y. (1996). A novel multicriteria optimization algorithm for the structure determination of multilayer feedforward neural networks. *Journal of Network and Computer Applications*, 19(2):135–147. 21

Krzywinski, M. and Altman, N. (2013). Points of significance: Significance, p values and t-tests. 82

Kurková, V. (2005). Neural network learning as an inverse problem. *Logic Journal of the IGPL*, 13(5):551–559. 31

Law, M. H., Topchy, A., and Jain, A. K. (2004). Multiobjective data clustering. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 424–430. 23

Lawson, C. L. and Hanson, J. R. (1974). *Solving Least Square Problems*. Prentice Hall, Englewoods Clifs, NJ. 35

Liu, C., Xu, X., and Hu, D. (2015). Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398. 23

Liu, G. P. and Kadirkamanathan, V. (1995). Learning with multi-objective criteria. In *Fourth International Conference on Artificial Neural Networks*, pages 53–58. IEEE. 21

Llorà, X. and Goldberg, D. E. (2003). Bounding the Effect of Noise in Multiobjective Learning Classifier Systems. *Evolutionary Computation*, 11(3):279–298. 23

Llorà, X., Goldberg, D. E., Traus, I., and Bernadó, E. (2002). Accuracy, parsimony, and generality in evolutionary learning systems via multiobjective selection. In *Learning Classifier Systems*, pages 118–142. Springer. Lecture Notes in Artificial Intelligence Vol. 2661. 22

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447. 24

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press. 14

Marler, R. T. and Arora, J. S. (2010). The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41(6):853–862. 45

Matsuyama, Y. (1996). Harmonic competition: a self-organizing multiple criteria optimization. *IEEE Transactions on Neural Networks*, 7(3):652–668. 21

Maulik, U., Bandyopadhyay, S., and Mukhopadhyay, A. (2011). *Multiobjective genetic algorithms for clustering: applications in data mining and bioinformatics*. Springer Science & Business Media. 23

McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133. 17

Medeiros, T. (2007). A new decision strategy in multi-objective training of artificial neural networks. pages 555–560. 6, 7, 28

Medeiros, T. H. (2006). Análise crítica do desempenho dos algoritmos multi-objetivo e por regularização bayesiana na resolução de problemas convexos e não-convexos. In *IX Encontro de Modelagem Computacional - EMC*, Belo Horizonte-MG. 45

Medeiros, T. H., Braga, A. P., and Takahashi, R. H. (2009). A incorporação do conhecimento prévio na estratégia de decisão do aprendizado multi-objetivo. In *Anais do Congresso Brasileiro de Redes Neurais*. 28

Medeiros, T. H., Rocha, H. P., Torres, F. S., Takahashi, R. H. C., and Braga, A. P. (2017). Multi-objective decision in machine learning. *Journal of Control, Automation and Electrical Systems*, 28(2):217–227. 6, 7, 28

Messac, A., Sundararaj, J. G., Tappeta, R. V., and Renaud, J. E. (2000). Ability of objective functions to generate points on non-convex pareto frontiers. *AIAA Journal*, 38(6):1084–1091. 120

Michie, D., Spiegelhalter, D. J., and Taylor, C. C., editors (1994). *Machine Learning, Neural and Statistical Classification*. Prentice Hall. 1

Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM*, 1:52–74. 35

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. 1, 9

Moffaert, K. V. and Nowe, A. (2014). Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research*, 15(1):3483–3512. 23

Mossalam, H., Assael, Y. M., Roijers, D. M., and Whiteson, S. (2016). Multi-objective deep reinforcement learning. 23

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., and Coello, C. A. C. (2013). Survey of multiobjective evolutionary algorithms for data mining: Part ii. *IEEE Transactions on Evolutionary Computation*, 18(1):20–35. 23

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., and Coello, C. A. C. (2014). A survey of multiobjective evolutionary algorithms for data mining: Part i. *IEEE Transactions on Evolutionary Computation*, 18(1):4–19. 23

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.]. 14

Neath, A. A. and Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, 4(2):199–203. 14

Oneto, L. (2018). Model selection and error estimation without the agonizing pain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8:e1252. 12

Pappa, G. L., Freitas, A. A., and Kaestner, C. A. A. (2004). Multi-objective algorithms for attribute selection in data mining. In Coello, C. A. C. and Lamont, G. B., editors, *Applications of Multi-Objective Evolutionary Algorithms*, pages 603–626. World Scientific. 23

Pareto, V. (1896). *Cours D'Economie Politique*, volume I and II. Rouse, Lausanne. 19

Parreiras, R. O. (2006). *Algoritmos Evolucionários e Técnicas de Tomada de Decisação em Análise Multicritério*. PhD thesis, Escola de Engenharia, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais. 6, 21, 26

Peebles, P. Z. (1993). *Probability, Random Variables, and Random Signal Principles*. McGraw-Hill, Singapore, 3rd ed. edition. 401 pages. 55, 91

Piironen, J. and Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735. 14

Ramlau, R. (2001). Morozov's discrepancy principle for tikhonov regularization of nonlinear operators. Technical report. 35

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808. 12, 15

Rissanen, J. (1978). Modeling by shortest data description*. *Automatica*, 14:465–471. 14

Rocha, H. P. (2017). *Treinamento Multiobjetivo de Perceptron de Múltiplas Camadas com a Representação Esférica de Pesos*. PhD thesis, Escola de Engenharia, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais. 25

Rocha, H. P., Costa, M. A., and Braga, A. P. (2015). Training multi-layer perceptron with multi-objective optimization and spherical weights representation. In *Proceedings of the European symposium on neural networks*, pages 131–136. 25

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536. 22

Sawaragi, Y., Nakayama, H., and Tanino, T. (1985). *Theory of Multiobjective Optimization*. Academic Press. 19

Schaffer, J. D. and Grefenstette, J. J. (1985). Multiobjective Learning via Genetic Algorithms. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 593–595, Los Angeles, California. AAAI. 21

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464. 14

Senhaji, K., Ramchoun, H., and Ettaouil, M. (2019). Multilayer perceptron: Nsga ii for a new multi-objective learning method for training and model complexity. In Mizera-Pietraszko, J., Pichappan, P., and Mohamed, L., editors, *Lecture Notes in Real-Time Intelligent Systems*, pages 154–167, Cham. Springer International Publishing. 23

Shaikhina, T. and Khovanova, N. A. (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*, 75:51 – 63. 6

Smith, C. and Jin, Y. (2014). Evolutionary multi-objective generation of recurrent neural network ensembles for time series prediction. *Neurocomputing*, 143:302 – 311. 24

Smola, A. J., Bartlett, P., Schölkopf, B., and Schuurmans, C. (1999). *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA. 12

Suttorp, T. and Igel, C. (2006). *Multi-Objective Optimization of Support Vector Machines*, pages 199–220. Springer Berlin Heidelberg, Berlin, Heidelberg. 5, 22

Teixeira, R. A. (2001). *Treinamento de Redes Neurais Artificiais Através de Otimização Multi-Objetivo: Uma Nova Abordagem para o Equilíbrio entre a Polarização e a Variância*. PhD thesis, Escola de Engenharia da UFMG. 22

Teixeira, R. A. (2002). Seleção de modelos neurais através de otimização multi-objetivo: Estratégias para implementação do decisor. *Proceedings of Seventh Brazilian Symposium on Neural Networks*, 1(1):112–118. 4, 6, 17, 22, 28, 38

Teixeira, R. A., Braga, A. P., Saldanha, R. R., Takahashi, R. H., and Medeiros, T. H. (2007). The usage of golden section in calculating the efficient solution in artificial neural networks training by multi-objective optimization. In *International Conference on Artificial Neural Networks*, pages 289–298. Springer. 6, 28

Teixeira, R. A., Braga, A. P., Takahashi, R. H. C., and Saldanha, R. R. (2000). Improving generalization of mlps with multi-objective optimization. *Neurocomputing*, 35(1):189–194. 3, 4, 5, 6, 22, 24, 28

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill Posed Problems*. Vh Winston. 11, 12, 32, 34

Torres, L. C., Castro, C. L., and Braga, A. P. (2012). A computational geometry approach for pareto-optimal selection of neural networks. In *International Conference on Artificial Neural Networks*, pages 100–107. Springer. 28

Torres, L. C. B. (2012). Uma nova abordagem baseada em margem para seleção de modelos neurais. Master's thesis, Escola de Engenharia, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais. 6

Triantaphyllou, E. (2000). Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study*, pages 5–21. Springer. 21

Vamplew, P., Dazeley, R., Berry, A., Issabekov, R., and Dekker, E. (2011). Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine learning*, 84(1-2):51–80. 23

Vamplew, P., Issabekov, R., Dazeley, R., Foale, C., Berry, A., Moore, T., and Creighton, D. (2017). Steering approaches to pareto-optimal multiobjective reinforcement learning. *Neurocomputing*, 263:26–38. 23

Van Moffaert, K., Drugan, M. M., and Nowé, A. (2013). Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199. 23

Vapnik, V. and Chervonenkis, A. Y. (1974). The method of ordered risk minimization, i. *Avtomatika i Telemekhanika*, 8:21–30. 3, 4, 12

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, NY. 736 pages. 10, 11

Vieira, D., Vasconcelos, J., and Saldanha, R. (2010). Recent advances in neural networks structural risk minimization based on multiobjective complexity control algorithms. In *Machine Learning*. IntechOpen. 22

Vito, E. D., Rosasco, L., Caponnetto, A., Giovannini, U. D., and Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904. 5, 31

Vogel, C. R. (1996). Non-convergence of the L-curve regularization parameter selection method. *Inverse Problems*, 12:535–547. 37

Weakliem, D. L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397. 14

Weigend, A. S., Rumelhart, D. E., and Huberman, B. A. (1990). Generalization by weight-elimination with application to forecasting. *Advances in Neural Information Processing Systems (NIPS 90)*, 3:879–882. 12

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edition. 14

Xu, Y. and Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3):249–262. 15

Yeung, D. S., Li, J., Ng, W. W. Y., and Chan, P. P. K. (2016). Mlpnn training via a multiobjective optimization of training error and stochastic sensitivity. *IEEE Transactions on Neural Networks and Learning Systems*, 27(5):978–992. 25

Zhang, J., Zhou, A., and Zhang, G. (2015a). A classification and pareto domination based multiobjective evolutionary algorithm. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 2883–2890. IEEE. 23

Zhang, J., Zhou, A., and Zhang, G. (2015b). A multiobjective evolutionary algorithm based on decomposition and preselection. In *Bio-inspired computing-theories and applications*, pages 631–642. Springer. 23

Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271. 22