UNIVERSIDADE ESTADUAL DE PONTA GROSSA
MESTRADO EM COMPUTAÇÃO APLICADA

ALISON ROGER HAJO WEBER

PROGRAMAÇÃO GENÉTICA, REDES NEURAIS ARTIFICIAIS E TÉCNICAS DE BALANCEAMENTO NA MODELAGEM DE DADOS AGRÍCOLAS: ESTUDO DA DOENÇA MOFO BRANCO

PONTA GROSSA
2012

ALISON ROGER HAJO WEBER

PROGRAMAÇÃO GENÉTICA, REDES NEURAIS ARTIFICIAIS E TÉCNICAS DE BALANCEAMENTO NA MODELAGEM DE DADOS AGRÍCOLAS: ESTUDO DA DOENÇA MOFO BRANCO

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Orientação: Dra. Aurora Trinidad Ramirez Pozo
Co-orientação: Dra. Alaine Margarete Guimarães

PONTA GROSSA
2012

Ficha Catalográfica Elaborada pelo Setor Tratamento da Informação Belém/UEPG

Weber, Alison Roger Hajo
W373 Programação genética, redes neurais artificiais e técnicas de balanceamento na modelagem de dados agrícolas : estudo da doença mofo branco / Alison Roger Hajo Weber . Ponta Grossa, 2012.
74 f.

Dissertação (Mestrado em Computação Aplicada), Universidade Estadual de Ponta Grossa.
Orientadora: Profa. Dra. Aurora Trinidad Ramirez Pozo.
Coorientadora: Profa. Dra Alaine Margarete Guimarães.

1. Modelagem. 2. Inteligência artificial. 3. Escleródio. 4. desbalanceamento. 5. Regressão. I. Pozo, Aurora Trinidad Ramirez II. Guimarães, Alaine Margarete. III. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. IV. T.

CDD: 006.3

TERMO DE APROVAÇÃO

ALISON ROGER HAJO WEBER

"PROGRAMAÇÃO GENÉTICA, REDES NEURAIS ARTIFICIAIS E TÉCNICAS DE BALANCEAMENTO NA MODELAGEM DE DADOS AGRÍCOLAS: ESTUDO DA DOENÇA MOFO BRANCO."

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora.

Orientador: Aurora Trinidad Ramirez Pozo
UFPR

David de Souza Jaccoud Filho
UEPG

Ivo Mario Mathias
UEPG

Ponta Grossa, 01 de Agosto de 2012.

AGRADECIMENTOS

Agradeço primeiramente a Deus que me ilumina dia-a-dia para atingir meus objetivos. Aos meus pais que de forma expressiva representam seu carinho e afeto nos momentos mais difíceis durante a caminhada. A Jessica, minha namorada, que também compreendeu entendeu a necessidade de minhas ausências e que também sempre me estimulou positivamente no sentido da importância deste trabalho. A todos meus familiares que na medida do possível podem contribuir com seu tempo de forma muito significativa, auxiliando e apoiando. As minhas orientadoras deste trabalho, que sem duvida alguma tiveram grande importância no desenvolvimento do mesmo. Aos meus colegas e professores do Mestrado, principalmente aos colegas Carlos, Cristian e Juscelino, pelos bons momentos que passamos juntos. A todo o pessoal do grupo de Fitopatologia da Universidade Estadual de Ponta Grossa, principalmente ao Professor Dr. David de Souza Jaccoud Filho e o Felipe Fadel Sartori, que com o apoio do Edital 064/2008 CNPq/MAPA contribuíram no desenvolvimento deste trabalho.

RESUMO

Problemas de regressão de dados são comuns na literatura, neles deseja-se inferir a relação entre variáveis dependentes (saída) e variáveis independentes (entrada) a partir de um conjunto de dados. Inferir esta relação entre as variáveis não é uma tarefa simples, por muitas vezes existirem uma alta não linearidade nos dados e pelo ruído existente neles. Duas técnicas de aprendizagem de máquina que são capazes de trabalhar com este tipo de informação são investigadas, a Programação Genética e as Redes Neurais Artificiais. Ainda assim, em muitos casos a técnica de aprendizado de máquina não consegue encontrar uma solução satisfatória, devido ao desbalanceamento da base de dados. Portanto, o objetivo deste trabalho foi aplicar técnicas de aprendizagem de máquina na regressão de dados desbalanceados, avaliando e comparando os resultados obtidos com diferentes abordagens. O método de balanceamento empregado resume-se em construir pesos para o conjunto de dados, um para cada exemplo, que representa a importância do exemplo durante o processo de aprendizagem do modelo. Este problema de modelagem em dados desbalanceados aplica-se em um caso real de modelagem de dados agronômicos, mais especificamente no estudo da doença mofo branco, causada pelo fungo *Sclerotinia sclerotiorum* (Lib.) de Bary. Devido ao alto poder destrutivo da doença para as culturas, o conhecimento da presença das estruturas de resistência chamadas de escleródios em uma área é de suma importância para que se tomem atitudes adequadas para o tratamento da doença. Neste estudo de caso, a tarefa é utilizar as técnicas de aprendizagem para a construção de um modelo de previsão de escleródios a partir de características meteorológicas e do local da amostra para o estado do Paraná, utilizando um conjunto de dados desbalanceados. Diferentes abordagens com as técnicas e com o método de balanceamento foram empregadas na construção do modelo. As Redes Neurais Artificiais com o algoritmo de aprendizagem *resilient propagation* obtiveram um melhor desempenho na criação do modelo para previsão de escleródios, conseguindo prever o resultado real com uma correlação de 0,763 e um erro médio absoluto de 24,35. Para identificar se o método de balanceamento empregado melhorou os resultados obtidos foi aplicado o teste de Kruskal-Wallis. O teste mostrou que existe uma melhora estatisticamente significativa entre a programação genética com e sem a técnica de balanceamento. Porem a técnica que apresentou melhores resultados foi a Rede Neural com o algoritmo de aprendizagem *resilient propagation*, no conjunto de dados do mofo branco e em alguns casos experimentais.

Palavras-chave: modelagem, escleródio, computação, regressão, desbalanceados.

ABSTRACT

Data regression problems are common in the literature, therein it is desired to infer the relationship between the dependent (output) and independent variable (input) from a dataset. Infer the relationship between variables is not a simple task, many times there is a high non-linearity and noise in the data inside them. Two machine learning techniques that are able to work with this type of information are investigated, the Genetic Programming and Artificial Neural Networks. Still, in many cases the machine learning technique cannot find a satisfactory solution due to the unbalance of the database. Therefore, the aim of this study was to apply machine learning techniques in regression of unbalanced data, evaluating and comparing the results obtained with different approaches. The balancing method used is summarized in constructing weights to the data set, one for each sample, which represents the importance of example during the learning process model. This problem of unbalanced data modeling applies in a real agronomic data modeling, specifically in the study of white mold disease caused by the fungus *Sclerotinia sclerotiorum* (Lib.) de Bary. Due to the high destructive power of the disease to crops, knowledge of the presence of resistance structures called sclerotia in an area is of paramount importance so that appropriate actions are taken to treat the disease. In this case study, the task is to use learning techniques to build a predictive model of sclerotia from meteorological characteristics and location of the sample to the state of Paraná, using a set of unbalanced data. Different approaches to the techniques and the balancing method was employed for constructing the model. The Artificial Neural Networks with resilient propagation learning algorithm achieved better performance in creating the model for prediction of sclerotia able to predict the actual outcome with a correlation of 0.763 and a mean absolute error of 24.35. To identify if the employee balancing method improved the results we applied the Kruskal-Wallis test. The test showed that there is a statistically significant improvement between genetic programming with and without balancing technique. However the technique that showed the best results was the neural network with resilient propagation learning algorithm, the data set of white mold and in some cases experimental.

Palavras-chave: modeling, sclerotia, computing, regression, unbalanced.

LISTA DE ILUSTRAÇÕES

Figura 1 - Estrutura Básica da Programação Genética. 15
Figura 2 - Árvore de Sintaxe da Programação Genética 17
Figura 3 - Exemplo de cruzamento..... 19
Figura 4 - Exemplo de mutação. 20
Figura 5 - Modelo matemático de um neurônio artificial. 22
Figura 6 - Representação de uma RNA. 23
Figura 7 - Ilustração simplificada do processo de aprendizado de uma RNA. 24
Figura 8 - Comparação entre os quatro métodos de construção de pesos apresentados..... 32
Figura 9 - Fluxograma da metodologia aplicada nos experimentos. 34
Figura 10 - Escleródios no solo. 44
Figura 11 - Ciclo de vida de *Sclerotinia sclerotiorum*..... 45
Figura 12 - Fluxograma da metodologia aplicada nos experimentos. 47
Figura 13 - Previsão do modelo utilizando PG com balanceamento. 55
Figura 14 - Previsão do modelo utilizando RNA com dados balanceados. 56

LISTA DE TABELAS

Tabela 1 - Informações sobre os conjuntos de dados experimentais..... 36
Tabela 2 - Resultados obtidos nos conjuntos de dados experimentais. 41
Tabela 3 - Saída do teste de Kruskal-Wallis para Correlação – bases experimentais. 41
Tabela 4 - Saída do teste de Kruskal-Wallis para RMSE – bases experimentais. 42
Tabela 5 - Amostra do conjunto de dados Mofo Branco..... 48
Tabela 6 - Conjunto de dados mofo branco normalizado. 49
Tabela 7 - Melhores configurações da PG. 51
Tabela 8 - Melhores configurações da RNA. 51
Tabela 9 - Medidas de desempenho dos modelos para previsão de escleródio. 54
Tabela 10 - Saída do teste de Kruskal-Wallis para Correlação – previsão de escleródios..... 54
Tabela 11 - Saída do teste de Kruskal-Wallis para RMSE – previsão de escleródios. 54
Tabela 12 - Saída do teste de Kruskal-Wallis para MAE – previsão de escleródios..... 54
Tabela 13 - Resultados obtidos no conjunto de dados de hardware de computador. 66
Tabela 14 - Resultados obtidos no conjunto de dados de resistência do concreto. 67
Tabela 15 - Resultados obtidos no conjunto de dados de incêndios florestais..... 68
Tabela 16 - Resultados obtidos no conjunto de dados de telemonitorização. 69
Tabela 17 - Resultados obtidos no conjunto de dados de qualidade do vinho. 70
Tabela 18 - Resultados obtidos no conjunto de dados do câncer de mama. 71
Tabela 19 - Resultados obtidos no conjunto de dados do mofo branco..... 72

LISTA DE SIGLAS

AG	Algoritmos Genéticos
AM	Aprendizagem de Máquina
CE	Computação Evolucionária
CPU	*Central Processing Unit*
GPS	*Global Positioning System*
IA	Inteligência Artificial
MAE	Erro Médio Absoluto
MSE	Erro Quadrático Médio
PAAF	Punção Aspirativa por Agulha Fina
PG	Programação Genética
RGP	R *Genetic Programming framework*
RMSE	Raiz do Erro Quadrático Médio
RNA	Redes Neurais Artificiais
Rprop	*Resilient Propagation*
UPDRS	*Unified Parkinson's Disease Rating Scale*

SUMÁRIO

1 INTRODUÇÃO 10
1.1 OBJETIVO 12
1.2 ORGANIZAÇÃO DO TRABALHO 13

2 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL 14
2.1 PROGRAMAÇÃO GENÉTICA 14
2.1.1 Representação de programas 16
2.1.2 Seleção de programas 18
2.1.3 Cruzamento e Mutação 19
2.1.4 Função de aptidão 20
2.1.5 Tarefa de regressão de dados 21
2.2 REDES NEURAIS ARTIFICIAIS 21
2.2.1 Introdução 21
2.2.2 Backpropagation 23
2.2.3 Resilient propagation 26
2.3 BALANCEAMENTO DE DADOS 28
2.3.1 Balanceamento de exemplos 29
2.3.2 Aplicando o balanceamento 31

3 EXPERIMENTOS E VALIDAÇÃO 34
3.1 CONJUNTOS DE DADOS 34
3.2 METODOLOGIA 36
3.2.1 Implementação 36
3.2.2 Seleção de parâmetros 37
3.2.3 Execução dos experimentos 39
3.3 RESULTADOS 41

4 ESTUDO DE UM CASO REAL: MOFO BRANCO 43
4.1 A DOENÇA MOFO BRANCO 43
4.1.1 Introdução 43
4.1.2 O ciclo da doença na planta 44
4.1.3 Modelagem da doença 46
4.2 METODOLOGIA 47
4.2.1 Conjunto de dados 48
4.2.2 Normalização 49
4.2.3 Parametrização das técnicas 50
4.2.4 Criação do modelo 52
4.3 RESULTADOS 52
4.3.1 Parâmetros das técnicas 52
4.3.2 Método de construção de pesos 53
4.3.3 Modelo para previsão de escleródios 54

5 CONCLUSÕES 58
5.1 TRABALHOS FUTUROS 58

REFERÊNCIAS 60

APÊNDICE A – Resultados dos trinta modelos em cada abordagem 65

1 INTRODUÇÃO

O objetivo de uma modelagem empírica é inferir dependências ocultas a partir de um conjunto de dados. Em um problema de regressão, a tarefa está em utilizar o conjunto de dados para representar variáveis de saída como funções analíticas de algumas, ou de todas as variáveis de entrada (FAYYAD et al., 1996). Inferir esta relação entre as variáveis não é uma tarefa simples, por muitas vezes existir uma alta não linearidade nos dados e pelo ruído existente neles, que podem ser definidos como exemplos que aparentemente são inconsistentes com o restante dos exemplos no conjunto de dados, pois não seguem o mesmo padrão que os demais (BARNETT; LEWIS, 1994).

Assim sendo, estudar e aplicar a técnica de aprendizagem adequada para inferir a relação existente nos dados é de grande relevância para obter um bom modelo (CHERKASSKY; MULIER, 2007). Dentro área de Inteligência Artificial existe uma subárea chamada de Aprendizagem de Máquina, dedicada ao desenvolvimento de algoritmos e técnicas que são capazes de aprender modelos (FAYYAD et al., 1996). Duas técnicas robustas, capazes de aprender modelos não linearmente separáveis e tolerantes a ruídos são a Programação Genética (PG) e as Redes Neurais Artificiais (RNA).

A PG é uma técnica que conseguem automatizar o processo de regressão baseada na teoria da evolução das espécies, na qual indivíduos são representados por programas de computador, com uma literatura rica em estudos de casos (KOZA, 1992; CHION et al., 2008; DABHI; VIJ, 2011). A RNA é uma técnica que busca simular o comportamento do cérebro humano, podendo esta ser aplicada em problemas de regressão, obtendo bons resultados em muitos estudos de caso (EL-TELBANY et al., 2006; JI et al., 2007). Apesar de bons resultados em diferentes domínios, existem casos onde a técnica não consegue encontrar uma solução satisfatória.

Em muitos problemas a técnica de modelagem não encontra uma relação satisfatória entre as entradas e saídas, o que pode significar que os dados utilizados no estudo não contenham todas as informações necessárias para predizer as saídas, este tipo de problema pode ocorrer pela falta de qualidade nos dados utilizados, que compromete a análise dos resultados obtidos (CHERKASSKY; MULIER, 2007).

Em um cenário ideal, os dados para a regressão compreendem observações de saída para quase todas as combinações possíveis de entradas. Considera-se então que os dados são equilibrados ou balanceados durante o processo das combinações, o que na maioria das vezes significa que os exemplos são distribuídos uniformemente sobre o espaço de entrada, e

consequentemente proporcionam uma quantidade semelhante de informações sobre sua respectiva resposta. Nesta situação o algoritmo de modelagem trata igualmente todas as amostras durante o processo de aprendizagem (CHERKASSKY; MULIER, 2007).

Porem, muitas vezes certas regiões do conjunto de dados estão representadas por uma abundância de observações, enquanto em outras regiões faltam representantes. Quando ocorre esta diferença no conjunto de dados, considera-se que os dados estão desequilibrados ou desbalanceados (VLADISLAVLEVA et al., 2010). Geralmente a construção de modelos com dados desbalanceados apresenta o risco de se obter soluções com um bom desempenho apenas em áreas com maior abundância de observações no conjunto de dados, prejudicando características importantes dos modelos, como a robustez e a capacidade de generalização (FAYYAD et al., 1996; CARDIE; NOWE, 1997; PROVOST; FAWCETT, 1997).

Em pesquisas Vladislavleva et al. (2010) desenvolveram um trabalho que apresenta uma maneira de tratar o desbalanceamento de dados na construção de modelos utilizando a técnica de Programação Genética, a qual é adaptada para utilizar um conjunto de pesos, um para cada exemplo, que representa a importância relativa do exemplo entre todos pertencentes ao conjunto de dados.

Os estudos anteriormente citados originaram da pesquisa aplicada e são de grande aplicação na área de agronomia, na qual é comum existir problemas onde o objetivo é realizar algum tipo de predição, ou então aprender um modelo, porem em muitos casos existe o problema de desbalanceamento de dados que dificulta a aprendizagem do modelo. A pesquisa em dados desbalanceados aplica-se em um caso real de modelagem de dados empíricos, no estudo de uma doença de culturas agrícolas conhecida como mofo branco.

Dentre os vários fatores que podem limitar a obtenção de altos rendimentos de culturas agrícolas estão às doenças. Existem aproximadamente quarenta doenças causadas por fungos, bactérias, nematoides e vírus já identificados no Brasil para a cultura da soja e as perdas anuais de produção devido à incidência de doenças pode variar de pequenas quantidades a 100% da produção (COSTA, 2009).

Nas últimas safras da cultura da soja a doença que vem chamando muito atenção é o mofo branco, causada pelo fungo *Sclerotinia sclerotiorum* (Lib.) de Bary, que pode atingir mais de 400 culturas. Na safra 2007/2008 vários estados brasileiros relataram problemas com esta doença, principalmente em Minas Gerais, onde a doença teve alta incidência nas áreas acima de 900 m (ZANETTI, 2009). Em regiões sudoeste, leste de Goiás e entorno ao Distrito Federal, as perdas de produção chegaram a 60% devido à doença (NUNES, 2009).

Neste cenário o estudo do mofo branco é um fator importante e necessário na adoção de medidas adequadas e preventivas no controle da incidência da doença na agricultura (JULIATTI; JULIATTI, 2010). Existem abordagens que buscam descobrir novos conhecimentos através de experimentos em campo, como pode ser visto em (SILVA; CAMPOS, 2007) e também existem abordagens que estudam esta doença a partir de dados coletados no campo e estudados por meio de modelos de regressão logística, como pode ser visto em (HARIKRISHNAN; DEL RÍO, 2008).

Estudos anteriores mostram que dados meteorológicos têm grande importância no aparecimento ou não do Mofo Branco na lavoura (HARIKRISHNAN; DEL RÍO, 2008; MILA; CARRIQUIRY; YANG, 2004). Nestes trabalhos foram utilizados métodos de regressão linear para construir modelos de previsão da doença, que obtiveram bons resultados. Entretanto, a construção de modelo de previsão não é uma tarefa simples e muitas vezes conduz a previsões pouco eficientes. Os métodos convencionais de previsão fornecem resultados precisos quando os dados estudados apresentam um comportamento linear, porém quando há um grau elevado de não linearidade, estes métodos passam a ser pouco eficientes (WEISS E KULIKOWSKI, 1991).

Devido ao alto poder destrutivo do mofo branco, o conhecimento da presença das estruturas de resistência chamadas de escleródios na área é de suma importância, pois é a partir dele que a doença pode atingir as culturas, podendo este, permanecer no solo por vários anos (SAGATA, 2010).

Assim sendo, prever a quantidade de escleródios presentes em uma área é fundamental na identificação do aparecimento da doença, pois mesmo com uma pequena quantidade de escleródios a doença pode ocorrer, não havendo ainda uma definição clara da proporção de escleródios para a incidência da doença. O problema deste tipo de dado é que durante a coleta das informações existem muitas áreas onde a presença do escleródio é baixa ou nenhuma e em poucas áreas a presença da estrutura é alta, caracterizando em uma informação desbalanceada.

1.1 OBJETIVO

O presente trabalho teve como objetivo aplicar técnicas de inteligência artificial na regressão de dados desbalanceados, avaliando e comparando os resultados obtidos com diferentes abordagens. Os objetivos específicos do trabalho estão descritos a seguir:

- Estudar e implementar as técnicas de PG e RNA para construção de modelos de regressão.
- Estudar e implementar um método para reduzir o efeito do desbalanceamento de dados.
- Avaliar se o balanceamento de dados por meio da utilização de pesos aprimora a construção de modelos para estudos de casos experimentais utilizando PG e RNA.
- Estudo dos dados de previsão de escleródios no estado do Paraná.
- Aplicar as diferentes abordagens no estudo de caso mofo branco para a construção de um modelo de previsão de escleródios.

1.2 ORGANIZAÇÃO DO TRABALHO

Este trabalho foi dividido em cinco capítulos, incluindo esta introdução. No capítulo 2 são apresentadas duas técnicas de Inteligência Artificial. A seção 2.1 apresenta a técnica de Programação Genética, como são representados os programas, cruzamento e mutação. A seção 2.2 descreve a técnica de Redes Neurais Artificiais, com os algoritmos de aprendizagem *backpropagation* e *resilient*. Na seção 2.3 está descrito a abordagem para balanceamento de dados. No capítulo 3 são descritos os experimentos que foram realizados para a validação das técnicas utilizadas. O capítulo 4 apresenta o estudo de caso da doença mofo branco. E, por fim, o capítulo 5 descreve as conclusões sobre o trabalho.

2 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) tem como objetivo principal o desenvolvimento de paradigmas ou algoritmos que requerem o poder computacional para realizar tarefas cognitivas, ou seja, tentar representar o comportamento humano por meio de modelos computacionais (SAGE, 1990). Um modelo de IA é capaz de realizar três tarefas: armazenar conhecimento, aplicar o conhecimento armazenado na resolução de problemas e adquirir novo conhecimento por meio de experiência (SAGE, 1990).

Dentro da IA existe uma subárea chamada de Aprendizagem de Máquina (AM), dedicada ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa (FAYYAD et al., 1996). Algumas partes da Aprendizagem de Máquina estão intimamente ligadas à Mineração de Dados (MD) e estatística, focada nas propriedades dos métodos, assim como em sua complexidade computacional (FAYYAD et al., 1996).

Aprender dados não é uma tarefa simples e para que se consiga aprender um modelo é necessário escolher uma técnica adequada. Dentre várias técnicas de AM existentes, foram selecionadas duas, a Programação Genética (PG) e as Redes Neurais Artificiais (RNA).

Ambas as técnicas são robustas a ruídos presentes nos dados e exigem pouco conhecimento sobre o domínio do problema, apresentando um paralelismo natural e sendo capaz de aprender conjuntos linearmente separáveis (POLI et al., 2008; BRAGA, 2000). A PG destaca-se ainda por produzir um modelo simbólico, ou seja, equações, enquanto a RNA produz um modelo composto de pesos que representam o conhecimento do modelo. Neste capítulo é apresentada uma revisão sobre estas duas técnicas de AM.

2.1 PROGRAMAÇÃO GENÉTICA

Considerando os estudos do ser humano em relação à evolução e a sobrevivência natural das espécies, surgiu a Computação Evolucionária (CE), que é uma das áreas da Inteligência Artificial inspirada na teoria Charles Darwin (1859).

A Computação Evolucionária compreende um conjunto de técnicas de busca e otimização inspiradas na evolução natural das espécies. É importante ressaltar que, a ideia da CE não é reproduzir certos fenômenos que acontecem na natureza, mas sim utilizar conceitos genéricos que estão presentes nesses fenômenos para a resolução de problemas computacionais.

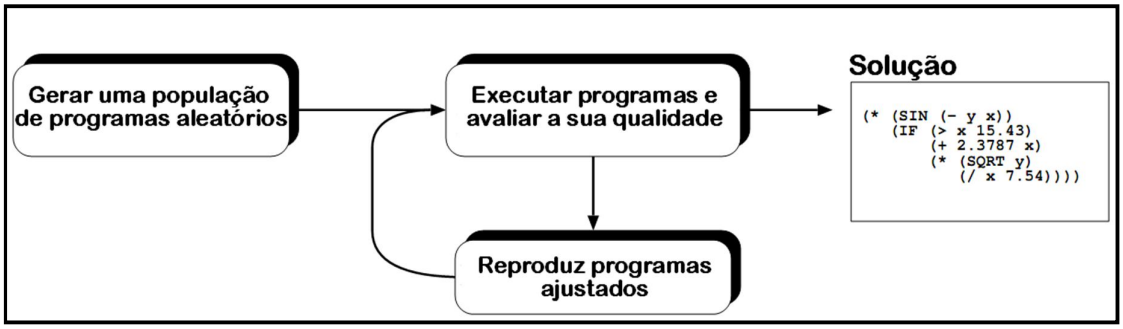
Segundo BANZHAF et al. (1998), a Computação Evolucionária abrange diversos ramos, entre eles estão os Algoritmos Genéticos, Estratégias de Evolução, Programação Evolucionária e a Programação Genética, esta última é abordada neste trabalho.

Com base nos trabalhos de John Holland em Algoritmos Genéticos (AG) (HOLLAND, 1975), a Programação Genética é uma técnica para a geração automática de programas de computador desenvolvida por John Koza (KOZA, 1989; 1992), baseada na combinação de ideias que surgiram ao longo do tempo, como a teoria da evolução com o conceito de seleção natural, operações genéticas (reprodução, cruzamento e mutação), a busca heurística vinda da Inteligência Artificial e a representação de programas como árvores sintáticas vinda da teoria de compiladores.

Na essência, a PG é um algoritmo que busca dentre um espaço relativamente grande, porém restrito de programas de computador, uma solução ou pelo menos uma boa aproximação para resolver determinado problema. Este é um algoritmo que pode não encontrar uma solução ótima, pois pode não varrer o espaço de busca completo, entretanto pode alcançar soluções em um menor tempo (POLI et al., 2008).

Na PG, o algoritmo evolutivo opera em uma população de programas computacionais que variam de forma e tamanho. Esta população de indivíduos será evoluída de modo a gerar uma nova população constituída por indivíduos com melhor aptidão, utilizando operadores de reprodução, cruzamento e mutação. Este processo é conduzido por uma função de aptidão, do termo em inglês *fitness*, que é utilizada como medida de quanto o indivíduo está próximo da solução do problema, sendo que os indivíduos possuidores de maior capacidade de adaptação têm maiores chances de sobreviver (KOZA, 1992). O funcionamento básico da PG pode ser visualizado na Figura 1, e o algoritmo de PG pode ser descrito resumidamente no algoritmo 1.

Figura 1 - Estrutura Básica da Programação Genética.



Fonte: Adaptado de Poli et. al., 2008.

Algoritmo 1: Programação Genética (Adaptado de Poli et. al., 2008, p. 3)

```
1: Gerar aleatoriamente uma população inicial de programas (Seção 3.3);
2: Repita
3:     Executar cada programa e avaliar sua aptidão por meio de uma função heurística (Seção 2.1.4);
4:     Enquanto número de filhos = tamanho da população
5:         Selecionar um ou dois programa(s) da população com base na aptidão para participar de operações genéticas (Seção 2.1.2);
6:         Criar novos programas pela aplicação de operações genéticas (Seção 2.1.3);
7:     Fim enquanto
8: Até que uma solução aceitável seja encontrada ou alguma outra condição de parada seja atendida (por exemplo: um número máximo de gerações);
9: Retornar o melhor indivíduo;
```

Ao iniciar o processo de aprendizagem com uma PG, o primeiro passo é efetuar a construção de uma população inicial de programas aleatoriamente de acordo com parâmetros definidos para a representação de programas, vide seção 3.3. Cada execução do laço **Repita** apresentado no Algoritmo 1 retorna uma nova população a cada geração, que é realizada avaliando cada programa por meio de uma função de aptidão, selecionando os melhores programas (Seção 3.4) e gerando novos por meio de operações genéticas (Seção 3.5).

O critério de parada para a geração de programas pode ser efetuado considerando diversas variáveis, mas comumente é estabelecido para atingir um número máximo de gerações (KOZA 1992). Também existem critérios de parada baseadas na análise do processo evolutivo, ou seja, a geração de populações permanece enquanto houver melhora nos indivíduos da população (KRAMER; ZHANG, 2000).

2.1.1 Representação de programas

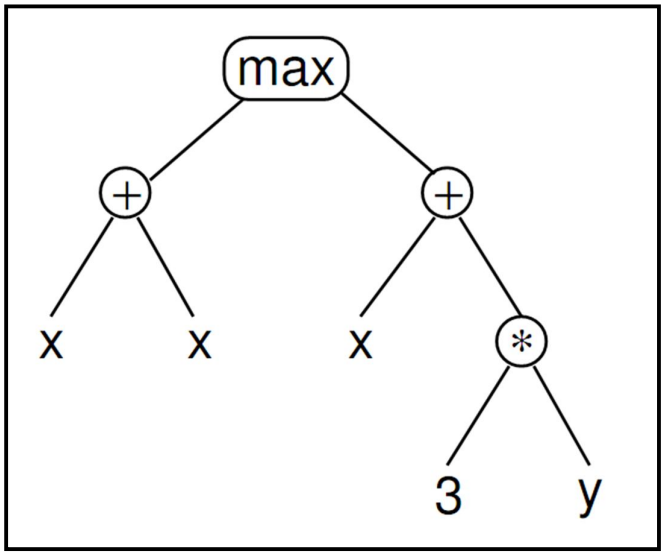
Para realizar a manipulação dos programas (indivíduos) a PG utiliza uma estrutura relativamente complexa e variável. Originalmente esta estrutura é uma árvore de sintaxe abstrata composta por funções em seus nós internos e terminais, os nós-folha, ao invés das

tradicionais linhas de código. A especificação do domínio do problema é feita simplesmente pela definição dos conjuntos de funções e terminais (KOZA, 1992).

Na representação de programas por árvores de sintaxe os individuos são formados por uma combinação dos conjuntos de Funções (F) e Terminais (T), de acordo com o domínio do problema. A exemplo disto tem-se um programa que possui a forma: max(x+x, x+3*y), sendo utilizado pela PG de acordo com a equação 1, com uma representação pré-fixa, semelhante ao utilizado em linguagens como Lisp ou Scheme (PEARCE, 1998). A representação deste indivíduo em forma de uma árvore de sintaxe é apresentada na Figura 2.

$$(\max\ (+\ x\ x)\ (+\ x\ (*\ 3\ y))) \qquad (1)$$

Figura 2 - Árvore de Sintaxe da Programação Genética



Fonte: Adaptado de Poli et. al., 2008.

Em todo algoritmo de Programação Genética deve-se definir inicialmente os conjuntos F de funções e T de terminais. No conjunto F, definem-se os operadores aritméticos, funções matemáticas, operadores lógicos, entre outros. O conjunto T é composto pelas variáveis e constantes e fornece um valor para o sistema (por exemplo, as variáveis de um conjunto de dados), enquanto o conjunto de funções processa os valores no sistema. Juntos, os conjuntos de funções e terminais representam os nós.

Pode-se citar como exemplo, o conjunto F, dos operadores aritméticos, e o conjunto T, de terminais, da seguinte forma:

$$F = \{ \max, +, * \} \text{ e } T = \{ x, y, 3 \}$$

Um indivíduo resultante da combinação dos conjuntos F e T pode ser o indivíduo apresentado na equação (1). A escolha dos conjuntos F e T influenciam consideravelmente, na solução apresentada pela PG. Se no conjunto F houver poucos operadores disponíveis, a PG provavelmente não será capaz de apresentar uma boa solução para o problema, porem ao disponibilizar muitos operadores o programa poderá ficar extenso, provocando esforço computacional desnecessário. O ideal é iniciar com operadores básicos: adição, subtração, multiplicação, divisão, conjunção, disjunção e negação e ir acrescentando outros operadores caso a solução apresentada não seja satisfatória.

Da mesma forma devem-se ter critérios para formar o conjunto das variáveis e constantes, pois o algoritmo de Programação Genética tem a capacidade de combinar as variáveis, transformando-as em novas variáveis (BANZHAF et al., 1998).

O espaço de busca da PG é constituído por todas as árvores que possam ser construídas por meio da combinação dos conjuntos F e T.

Em formas avançadas de PG, os programas podem ser compostos por múltiplos componentes (subrotinas). Neste caso, a representação utilizada na PG é um conjunto de árvores agrupadas sob um nó raiz especial, atuando como uma junção, nomeando esses ramos de subárvores, considerando o número e tipo de ramos em um programa, juntamente com algumas outras características de sua estrutura, formando a arquitetura do programa (Poli et. al., 2008).

2.1.2 Seleção de programas

Na diversidade dos algoritmos evolucionários, os operadores genéticos da PG são aplicados aos indivíduos que são probabilisticamente selecionados com base na aptidão (*fitness*), ou seja, os melhores indivíduos são mais aptos a terem um número maior de filhos, ao contrario dos indivíduos com um menor valor de aptidão.

Para realizar a seleção dos indivíduos existem diversos métodos que podem ser utilizados, dentre eles existem o método de seleção por roleta, *ranking*, truncamento e torneio (GOLDBERG, 1989).

No método de seleção por torneio a competição entre os indivíduos não é realizada em toda população, mas apenas num subconjunto da população. Certo número de indivíduos é selecionado, que é o tamanho do torneio, e uma competição seletiva é realizada. As características dos melhores indivíduos no torneio são substituídas pelas características dos

piores indivíduos. No menor torneio possível é permitido que dois indivíduos participem da reprodução (GOLDBERG, 1989).

O método do torneio é mais comumente utilizado, pois oferece a vantagem de não exigir que a comparação seja feita entre todos os indivíduos da população (BANZHAF et al., 1998).

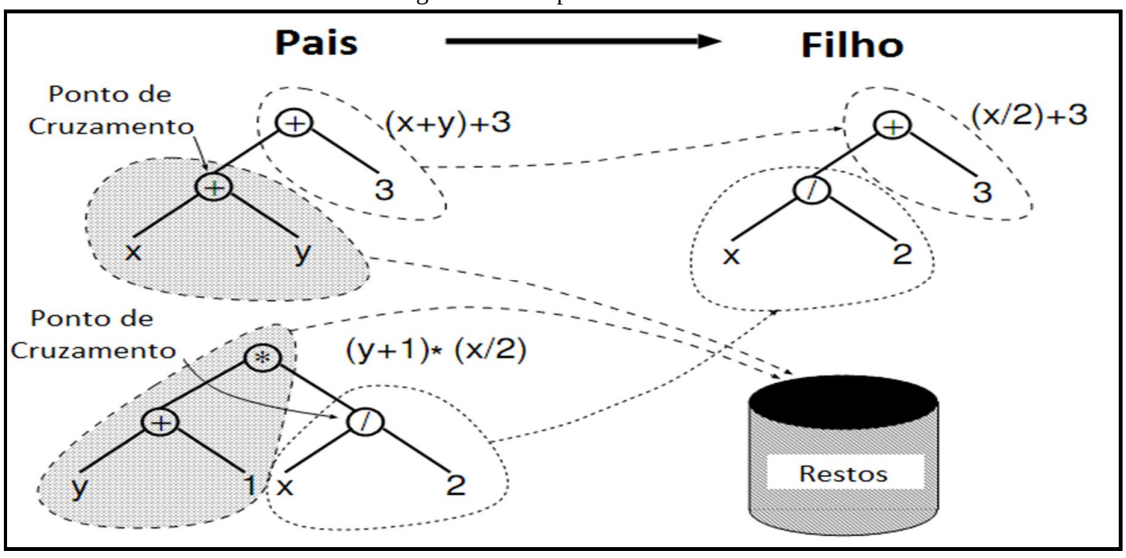
2.1.3 Cruzamento e Mutação

A PG afasta-se significativamente de outros algoritmos evolucionários na implementação dos operadores de cruzamento e mutação. A forma mais comum utilizada de cruzamento é o cruzamento subárvore.

O operador de cruzamento visa guiar a solução de maneira a combinar as melhores soluções na busca da solução ótima. Dado dois pais, o mecanismo seleciona um ponto de cruzamento (um nó) em cada árvore-pai. Então, ele cria os filhos substituindo a subárvore enraizada no ponto de cruzamento em uma cópia do primeiro pai com uma cópia da subárvore enraizada no ponto de cruzamento no segundo pai, como ilustrado na Figura 3. As cópias são utilizadas para impedir a alteração dos indivíduos originais, desta forma, se selecionado várias vezes, eles podem tomar parte na criação de programas de múltiplos filhos (KOZA, 1992).

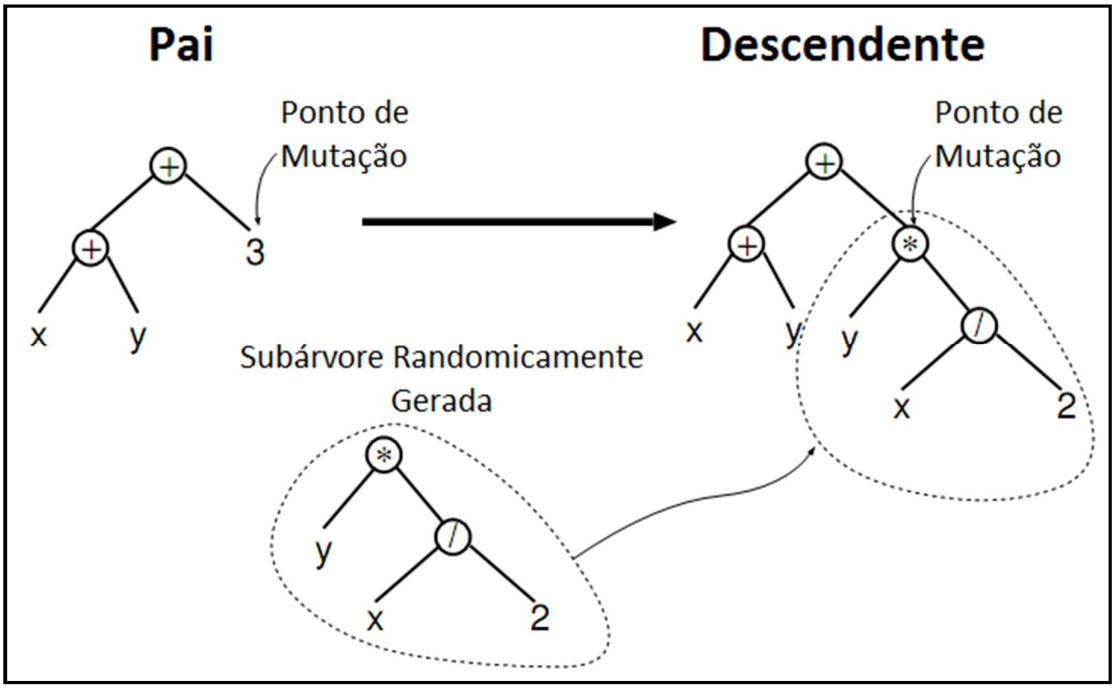
A forma comumente utilizada na mutação da PG seleciona aleatoriamente um ponto de mutação em uma árvore, e substitui a subárvore que foi enraizada naquele ponto, com uma subárvore gerada aleatoriamente. Isto é ilustrado na Figura 4. Muitos outros tipos de cruzamento e mutação das árvores PG são possíveis, vide Poli et. al. (2008).

Figura 3 - Exemplo de cruzamento.



Fonte: Adaptado de Poli et. al. (2008).

Figura 4 - Exemplo de mutação.



Fonte: Adaptado de Poli et. al. (2008).

2.1.4 Função de aptidão

Para definir uma função de aptidão é necessário conhecer o domínio do problema. Em geral, nos problemas de otimização esta função é definida como sendo a função objetivo, porém nada impede que se defina outra função. Escolher bem a função de aptidão pode resultar em um bom funcionamento da PG.

Especificamente, no caso de um problema de regressão, pode-se utilizar como função de aptidão, a função que mede o erro calculado entre o valor previsto e o valor real, como por exemplo, o erro quadrático médio. Quanto menor for o erro obtido, melhor será o ajuste do modelo de previsão. O que se deseja, portanto, é minimizar a função de aptidão (BANZHAF, 1998).

A função de aptidão é uma forma de se diferenciar os melhores dos piores indivíduos. Se esta função for bem definida há uma grande probabilidade de que o algoritmo gere uma solução muito próxima da solução ótima (KOZA, 1992).

2.1.5 Tarefa de regressão de dados

Uma das formas de encontrar programas de computador que resolvam um dado problema computacional é por meio da regressão simbólica, que consiste na manipulação de expressões matemáticas para descoberta de funções que representem um conjunto de dados. A atualmente existe várias técnicas computacionais capazes de automatizar a regressão simbólica. Uma dessas técnicas é a programação genética.

A regressão simbólica tem sido testada desde as primeiras implementações de programação genética e a literatura da área é rica em estudos de caso para problemas de modelagem matemática (KOZA, 1992).

2.2 REDES NEURAIS ARTIFICIAIS

2.2.1 Introdução

A busca por modelos computacionais que conseguissem simular o funcionamento de cérebro humano começa em 1943, com o trabalho McCulloch e Pitts, no qual foi desenvolvido um modelo simples de um neurônio artificial. Segundo Ferneda (2006) a motivação pela pesquisa nessa área cresceu muito durante os anos 50 e 60. Em 1958 Rosenblatt propôs um método inovador de aprendizagem para a RNA denominado *perceptron*. Até 1969 muitos trabalhos foram desenvolvidos utilizando o perceptron. Ainda em 1969 os pesquisadores Minsky e Pappert publicaram um livro apresentando importantes limitações do modelo *perceptron*, que solucionava apenas problemas que fossem linearmente separáveis. Com as limitações metodológicas e tecnológicas da época, juntamente com as criticas destes autores, fizeram com que os trabalhos e pesquisas na área fossem enfraquecidas. Muitos anos se passaram e as pesquisas não avançaram, porém, durante os anos 80, o entusiasmo volta graças aos avanços metodológicos, proposto com o modelo do *multilayer perceptron* e também ao aumento do poder computacional das máquinas.

Atualmente as Redes Neurais Artificiais (RNA) também referidas na literatura como Redes Neurais (RN), são definidas de diferentes formas por diferentes autores. Apesar de não haver uma única definição os autores, Haykin (2001), Braga e Ludermir (2000) concordam que as redes neurais são um método de solucionar problemas de Inteligência Artificial (IA), a partir de um sistema composto de neurônios e conexões, cuja finalidade é simular de forma

computacional o funcionamento do cérebro humano, ou seja, errando, aprendendo, e fazendo descobertas.

O modelo do neurônio artificial, vide Figura 5, é o modelo apresentado por Haykin (2001). Este modelo é composto por três elementos:

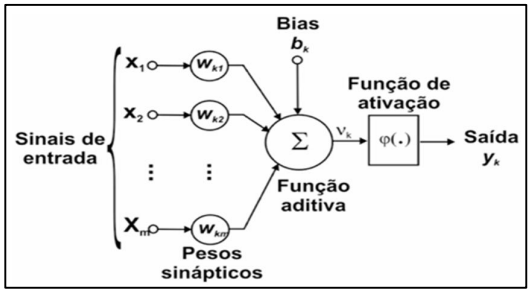
- Um conjunto de m conexões de entrada (x_1, x_2, ..., x_m) caracterizadas por um peso ou força própria (w_{k1}, w_{k2}, ..., w_{km}).
- Um somador para acumular os sinais de entrada, que pode ser representado pelas equações 2 e 3.

$$u_k = \sum_{j=i}^{m} w_{kj} x_i \tag{2}$$

$$v_k = u_k + b_k \tag{3}$$

- Uma função de ativação (φ) responsável pela ativação da saída (y) do neurônio artificial. Os três tipos básicos de função de ativação são: Função de Limiar, Função Linear por Partes e Função Sigmoide.

Figura 5 - Modelo matemático de um neurônio artificial.

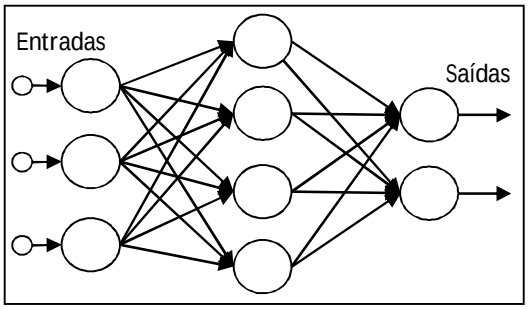


Fonte: Haykin, 2001.

O comportamento das conexões entre os neurônios é obtido por meio de seus pesos, que podem ser positivos ou negativos, dependendo do tipo da conexão, inibitória ou excitatória. O efeito do sinal recebido por outro neurônio é determinado pela multiplicação do valor do sinal pelo peso da conexão ($x_m * w_{km}$). Então é realizado o somatório de todos os valores das conexões, assim o valor resultante é enviado à função de ativação que define a saída (y).

Quando combinado diversos neurônios, é formada uma RNA, podendo ser representada como um grafo orientado, em que os nós são os neurônios e as ligações representam as sinapses, vide Figura 6.

Figura 6 - Representação de uma RNA.



Fonte: Haykin, 2001.

Haykin (2001) afirma que as redes neurais se diferenciam pela sua arquitetura e pela forma com que os pesos sinápticos são ajustados durante o processo de treinamento. Geralmente a arquitetura de uma rede neural é restringida para o problema ao qual ela foi construída. A arquitetura de uma rede é definida pelo seu número de camadas: única ou múltiplas, pelo número de neurônios em cada camada da rede, por sua topologia e pelo tipo de conexão entre os nós: *feedforward* (acíclica) ou *feedback* (cíclica).

Braga e Ludemir (2000) definem que uma das propriedades mais importante de uma rede neural artificial é a capacidade de aprender através de instancias e realizar inferências sobre o que aprendeu desta forma podendo melhorar o seu desempenho a cada nova iteração.

Existem diferentes formas de aprendizado em redes neurais, uma delas é o aprendizado supervisionado, no qual o método é responsável por apresentar conjuntos de padrões de entrada a rede e suas respectivas saídas, então para cada entrada, o método indica quanto houve de erro para resposta calculada. O erro encontrado é informado à rede neural para que seja feito os ajustes nos pesos sinápticos. Neste método é necessário que haja um conhecimento prévio do comportamento que se deseja ou que se espera da rede.

Existem vários algoritmos de aprendizado supervisionado que diferem entre si pela forma como é ajustado o peso sináptico, mas para este trabalho serão abordados apenas dois algoritmos, o *backpropagation* e o *resilient propagation.*

2.2.2 Backpropagation

Conforme Mueller (1996) por muitos anos não se teve um algoritmo de treinamento eficiente para redes neurais artificiais de múltiplas camadas. Desde que as redes neurais de única camada se mostraram limitadas em problemas que poderiam representar, e portanto, no

que poderiam aprender, o desenvolvimento de modelos de redes neurais deixou de ser interessante e poucos trabalhos foram realizados.

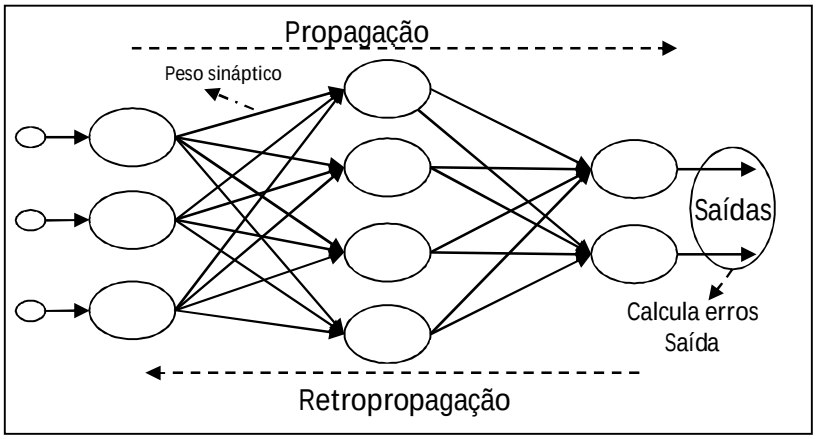
O algoritmo de aprendizagem supervisionada *backpropagation*, proposto por Werbos, Parker e Rummelhart, fez ressurgir o interesse em RNA, podendo ser visto com uma generalização do método Delta proposto por Widrow e Hoff.

Ao apresentar-se um padrão de entrada a uma RNA não treinada e seu respectivo padrão de saída a rede produzirá uma saída aleatória. A partir da saída calculada pela rede é calculado um erro, que representa a diferença entre o valor obtido e o esperado. Então o funcionamento deste algoritmo consiste em, reduzir continuamente o erro enquanto as condições de parada não forem satisfeitas. Este resultado é obtido pelo ajuste dos pesos entre as conexões dos neurônios, conhecidos como pesos sinápticos, aplicando a regra Delta Generalizado (HAYKIN, 2001).

Para atingir um bom resultado é necessário minimizar o erro do algoritmo, obtendo o gradiente descendente da função de erro. Desta forma o ajuste dos pesos inicia na camada de saída, onde a medida do erro está disponível e continua com a retropropagação desse erro entre as camadas, ajustando os pesos em todas as sinapses da rede. Na camada de saída, como os valores desejados e obtidos são conhecidos, o ajuste dos pesos sinápticos é simples, mas para as unidades presentes nas camadas ocultas este processo não é tão trivial. Comumente nas primeiras épocas de treinamento as camadas ocultas apresentam grandes erros e devem ter seus pesos alterados em grande escala.

Sendo assim é possível identificar duas fases distintas no processo de aprendizagem do *backpropagation*, na primeira as entradas se propagam até a saída e na segunda é quando os erros são retropropagados da camada de saída até a entrada (Figura 7).

Figura 7 - Ilustração simplificada do processo de aprendizado de uma RNA.



Fonte: Haykin, 2001.

Mueller (1996) apresenta as etapas do algoritmo *backpropagation*. Seja um conjunto de dados de treinamento (D), cada instancia deste conjunto é separada em dois vetores, um vetor de entrada (d_i) e um vetor com os dados da saída (d_o). O objetivo do treinamento é ensinar a rede o mapeamento de todo o vetor de entrada para o respectivo vetor de saída, isto é, estabelecer $x_0(w, d_i) = d_o$, pela configuração de valores adequada para cada peso (w) da rede.

O ajuste dos pesos da RNA é realizado sempre que a saída desejada e a saída real (x_0) não forem iguais, então é aplicado o método do gradiente descendente, equação 4.

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{4}$$

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{5}$$

Na equação 5, (η) é a taxa de aprendizagem e a derivada parcial da função de erro demonstrada na equação 6, é a medida do erro.

$$E = \frac{1}{2} \sum_p \sum_j (d_{oj} - x_{0j})^2 \tag{6}$$

Onde (d_{oj}) é o j-ésimo componente do vetor de saída. As etapas do *backpropagation* podem ser sumarizadas no algoritmo 2:

Algoritmo 2: Aprendizado de uma RNA.

```
1: Inicializar o conjunto de pesos (w) com valores aleatórios;
2: Repita
3:      Apresentar vetores de entrada e saída;
4:      Calcular resposta da rede;
5:      Calcular erro para as unidades de saída;
6:      Calcular erro para as unidades ocultas;
7:      Atualizar o peso de todas as camadas;
8: Até que uma solução aceitável seja encontrada ou enquanto um
   número determinado de iterações não for satisfeito;
9: Retornar rede neural treinada com os últimos pesos obtidos;
```

O *backpropagation* é um algoritmo de aprendizagem eficiente e utilizado em diversos trabalhos, mas sua implementação é trabalhosa, exigindo muitos passos e deixando o treinamento longo. Existem muitos estudos a respeito de melhorias a esta técnica, para que o algoritmo chegue a uma boa taxa de erro em menos tempo de treinamento e uma delas é o algoritmo *resilient propagation*.

2.2.3 Resilient propagation

Muitos algoritmos foram desenvolvidos para lidar com o problema de calcular o valor apropriado dos pesos das sinapses. Em 1993 Riedmiller e Braun propuseram um novo algoritmo de aprendizagem supervisionada de RNA multicamadas, chamado de *resilient propagation* (Rprop). Este é uma evolução do *backpropagation*, com o objetivo de eliminar o problema da "adaptabilidade desfocada", ou seja, para que a rede não de passos pequenos quando a derivada for pequena, o que deixará o processo de treinamento lento, assim o Rprop altera o valor da atualização do peso (Δw_{ij}) diretamente sem considerar o valor da derivada parcial, assim convergindo mais rápido.

O Rprop efetua uma adaptação direta do peso de cada passo baseado na informação do gradiente local, não sendo esta adaptação influenciada pelo valor do gradiente. Para que o algoritmo consiga realizar tal tarefa é necessário que cada sinapse possua além de seu peso, um valor de atualização, que será chamado de Δ_{ij}, o qual será utilizado para determinar o tamanho da atualização do peso. Este valor irá evoluir durante o processo de aprendizagem baseado na função de erro E, seguindo a regra de aprendizagem demonstrada na equação 7.

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} * \Delta_{ij}^{(t-1)}, se\ \dfrac{\partial E}{\partial \omega_{ij}}(t-1) * \dfrac{\partial E}{\partial \omega_{ij}}(t) > 0 \\ \eta^{-} * \Delta_{ij}^{(t-1)}, se\ \dfrac{\partial E}{\partial \omega_{ij}}(t-1) * \dfrac{\partial E}{\partial \omega_{ij}}(t) < 0 \\ \Delta_{ij}^{(t-1)}, se\ \dfrac{\partial E}{\partial \omega_{ij}}(t-1) * \dfrac{\partial E}{\partial \omega_{ij}}(t) = 0 \end{cases} \quad (7)$$

Se em um determinado momento o peso sináptico correspondente da derivada parcial w_{ij} inverter seu sinal, é porque sua ultima modificação foi elevada e o algoritmo saiu do mínimo local, então o tamanho da modificação deverá diminuir de acordo com o fator η^{-} e o valor prévio do coeficiente do peso sináptico é restabelecido. Em outras palavras, esta modificação deveria ser desfeita, como mostra a equação 8.

$$\Delta w_{ij}^{t} = \Delta w_{ij}^{t} - \Delta_{ij}^{(t-1)} \tag{8}$$

Se o sinal da derivada não foi alterado, então o ajuste deverá ser aumentado por η^{+} para alcançar mais rapidamente a convergência. Para evitar valores de pesos sinápticos muito altos ou muito baixos, o valor de ajuste é determinado acima por $\Delta_{\max}$ e abaixo por $\Delta_{\min}$.

Então quando o valor de atualização estiver adaptado, a atualização do peso segue a seguinte regra: se a derivada é positiva, o erro aumentou e o peso é diminuído pelo valor de atualização ou se a derivada é negativa o valor de atualização é somado, equação 9. A equação 10 apresentada a forma utilizada para encontrar o valor dos pesos sinápticos.

$$\Delta_{\omega_{ij}}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)}, se \dfrac{\partial E}{\partial \omega_{ij}}(t) > 0 \\ +\Delta_{ij}^{(t)}, se \dfrac{\partial E}{\partial \omega_{ij}}(t) < 0 \\ 0, se \dfrac{\partial E}{\partial \omega_{ij}}(t) = 0 \end{cases} \tag{9}$$

$$w_{ij}^{t+1} = w_{ij}^{t} + \Delta w_{ij}^{(t)} \tag{10}$$

O processo de treinamento utilizado no algoritmo Rprop segue as mesmas etapas do algoritmo 2. O Rprop necessita de alguns parâmetros e os autores Riedmiller e Braun (1993) sugerem a seguinte distribuição, que geralmente resulta em um bom tempo de convergência.

- Delta inicial. $\Delta_0 = 0.1$
- Delta Máximo. $\Delta_{\max} = 50.0$
- Delta Mínimo. $\Delta_{\min} = 1e^{-6}$
- Fator de redução. $\eta^{-} = 0.5$
- Fator de incremento. $\eta^{+} = 1.2$

2.3 BALANCEAMENTO DE DADOS

A necessidade de dados balanceados para a regressão na análise de dados origina-se de um campo de pesquisa aplicada. O mercado exige que a previsão de dados seja precisa e o desafio dos pesquisadores em aplicar e desenvolver modelos de previsão com dados imperfeitos é grande. Pensando nisso, Vladislavleva et al. (2010) desenvolveram um trabalho focado na importância do balanceamento dos dados para construção de modelos aplicado à PG.

Existe um conceito na estatística chamado de *outlier*, que é uma observação do conjunto de dados que apresenta um grande afastamento das demais observações, ou que é inconsistente. A existência de *outliers* implica em prejuízos a interpretação de resultados dos testes estatísticos aplicados ao conjunto de dados e a maioria de trabalhos de pesquisa voltada à análise de dados para modelagem computacional está direcionada principalmente na detecção de *outliers*, como pode ser visto em (AGGARWAL E YU, 2001; DIEDERIK et. al., 2003; MOTULSKY E BROWN, 2006).

Entretanto Vladislavleva et al. (2010) observa que a remoção dos *outliers* da base, pode não representar a melhor solução para o problema. Supondo que os dados disponíveis para modelagem são confiáveis, mas possivelmente possuem algum tipo de ruído, os *outliers* devem ser considerados como amostras contendo informações importantes sobre a saída. Sobre esta alternativa o objetivo não é detectar os *outliers* a fim de excluí-los, mas sim tratá-los com critério durante o processo de aprendizagem. Neste sentido os autores reproduzem evidencias e sugestões para somar e aprimorar o processo de regressão para conjunto de dados desbalanceados utilizando PG.

A sugestão proposta por Vladislavleva et al. (2010) é construir um conjunto de pesos (***w***), um para cada amostra do conjunto de dados, que representa a importância relativa do ponto entre as demais amostras. Este peso (*w*) pode ser definido por quatro métodos, que se diferenciam na maneira de calcular o peso, obtendo um melhor desempenho de acordo com a distribuição do dado que está sendo processado. Os quatro métodos são os seguintes, onde *k* é um parâmetro para definição de quantos vizinhos serão considerados no calculo do *w*:

1) Pela proximidade de *k* vizinhos em um espaço de vizinhos;
2) Pela distância circundante de *k* vizinhos em um espaço de vizinhos;
3) Pelo afastamento de *k* vizinhos em um espaço de vizinhos;

4) Pelo desvio da mais próxima entrada de uma passagem em um hiperplano com mínimos quadrados através de *k* vizinhos, para $k \geq d$, onde *d* é a dimensionalidade do espaço de entrada.

O valor de *k* é um valor que pode ser ajustado empiricamente, mas que varia de acordo com o tamanho do conjunto de dados e a precisão que se quer chegar. Para se chegar ao conjunto de pesos $\boldsymbol{w}$, primeiramente deve-se levar em consideração alguns aspectos essenciais para o procedimento.

Para definir quais exemplos são vizinhos um do outro é necessário estabelecer uma métrica de cálculo de distâncias e ao olhar para esquemas existentes procurando por um método adequado e escalável com base na distância de ponderação para dados multidimensionais, deve-se iniciar por ter uma distância métrica escalável. Ao utilizar, por exemplo, a métrica da distancia euclidiana, equação 11, onde p e q são dois pontos em um espaço, existe a dificuldade de obter uma noção significativa de proximidade em um espaço tridimensional:

$$L_2(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{i=1}^{d} (p_i - q_i)^2 \right)^{1/2} \tag{11}$$

Para solucionar o problema, principalmente em casos de alta dimensionalidade, é adotada uma métrica inspirada em AGGARWAL et al. (2001), equação 12, onde $\boldsymbol{u}$ e $\boldsymbol{v}$ são dois pontos no espaço:

$$dist_{\frac{1}{d}}(\boldsymbol{u}, \boldsymbol{v}) = \left(\sum_{i=1}^{d} |u^i - v^i|^{1/d} \right)^d \tag{12}$$

2.3.1 Balanceamento de exemplos

Vladislavleva et al. (2010) revisaram alguns métodos para ajudar no balanceamento de exemplos inserindo pesos sobre cada exemplo do conjunto de dados. Baseados em métodos de detecção de *outliers* propostos por Harmeling et al. (2006) os autores levantaram questões de como e até quando os dados desbalanceados são realmente relevantes, e também qual o conjunto de informações de entrada que é útil para o balanceamento do exemplo, apenas as entradas ou as entradas e a saída do exemplo. Durante estes estudos os autores verificaram que utilizando as entradas e saída do exemplo para determinar um peso de uma amostra existe uma melhora significativa nos resultados. Os métodos propostos por

Vladislavleva et. al. (2010) para determinação de pesos para as amostras são descritas a seguir.

- **Pela Proximidade**

Neste método o peso para a amostra é definido pela proximidade dos k vizinhos mais próximos utilizando as informações de entrada e saída para calculo da distancia, equação 13.

$$\pi(i, \mathcal{M}, k) = \frac{1}{k}\sum_{j=1}^{k} ||M_i - \bar{n}_j(M_i, \mathcal{M})|| \quad (13)$$

Onde $\bar{n}_j(M_i, \mathcal{M})$ é o j-ésimo vizinho mais próximo em um espaço de vizinhos da amostra M_i em um conjunto de amostras $\mathcal{M}$, .

- **Pela Distância Circundante**

Utilizando a mesma ideia de determinar a vizinhança no espaço de entrada esta outra forma de determinar pesos considera os vizinhos mais próximos de forma circundante, equação 14.

$$\sigma(i, \mathcal{M}, k) = ||\frac{1}{k}\sum_{j=1}^{k} (M_i - \bar{n}_j(M_i, \mathcal{M})|| \quad (14)$$

Onde $\bar{n}_j(M_i, \mathcal{M})$ é o j-ésimo vizinho mais próximo em um espaço de vizinhos da amostra M_i em um conjunto de amostras $\mathcal{M}$.

- **Pelo Afastamento**

Utilizando as características da proximidade e dos pesos circundantes introduzidos anteriormente apresenta-se outro método de estipulação de pesos através do afastamento das amostras. O objetivo deste método é atribuir peso maior para os pontos com maior teor de informação em relação aos vizinhos. Pesos elevados destes pontos irão conduzir o sistema a gerar menores erros em locais específicos. Portanto, pode-se atribuir pesos maiores para os pontos de borda para restringir as soluções a apresentar um comportamento de extrapolação mais precisa, atribuindo maior peso aos pontos nas regiões esparsas, forçando soluções para concordar com os dados de treinamento em áreas sobre representadas.

Segundo Vladislavleva et al. (2010) este método é melhor que os dois anteriores e resulta em modelos melhores na maioria de problemas, uma vez que combina o ranking de exemplos, e não os valores de peso reais. Este método de construção de pesos utiliza os métodos citados nas seções anteriores, conforme pode ser visto na equação 15.

$$\rho(i, \mathcal{M}, k) = \mathcal{L}[i; \mathcal{L}[i; \pi(i; \mathcal{M}, k)] + \mathcal{L}[i; \sigma(i, \mathcal{M}, k)]] \qquad (15)$$

Seja $\omega : \mathcal{M} \mapsto \omega(\mathcal{M}, k) \in \mathbb{R}$ um peso do conjunto de amostras $\mathcal{M} = \{M_1, M_2, \dots, M_N\} in\, \mathbb{R}^{d+1}$. O vetor de pesos $(\omega(1, \mathcal{M}, k), \dots, \omega(N, \mathcal{M}, k))$ induz a um ranking de valores de *w*. Seja $\mathcal{L}[i; \omega(\mathcal{M}, k)] \in \{1, \dots, N\}$ é um índice do valor $\omega(i, \mathcal{M}, k)$ no ranking de pesos.

Por um lado, o peso por afastamento $\rho(\cdot)$ pode detectar com sucesso pontos de alta variação na resposta em relação aos vizinhos mais próximos, porem, enfatizar os pontos de grandes mudanças na resposta com o índice de afastamento pode ser uma maneira de capturar as áreas de inclinação da superfície de resposta, uma tarefa não óbvia em conjuntos de dados com alta dimensionalidade. Por outro lado, o excesso de foco em pontos de alta variação na saída pode retirar a atenção das áreas de variação inferior, mas ainda altamente significativas.

- **Pelo Desvio do Mais Próximo em um Hiperplano**

O peso com base no desvio local a partir de linearidade da amostra de $M_i \in \mathbb{R}$ é definido como a distância da amostra M_0 para o hiperplano de mínimos quadrados que passa através do *k* vizinho mais próximo, onde $k \geq d + 1$, equação 16.

$$\mathcal{V}(i, \mathcal{M}, k) = dist_{XY}(M_i, \Pi_i) \qquad (16)$$

Onde Π_i é uma passagem pelo (k - 1) d-dimensional hiperplano do k vizinho mais próximo M_i e $dist_{XY}$ é a distância métrica selecionada do espaço M.

2.3.2 Aplicando o balanceamento

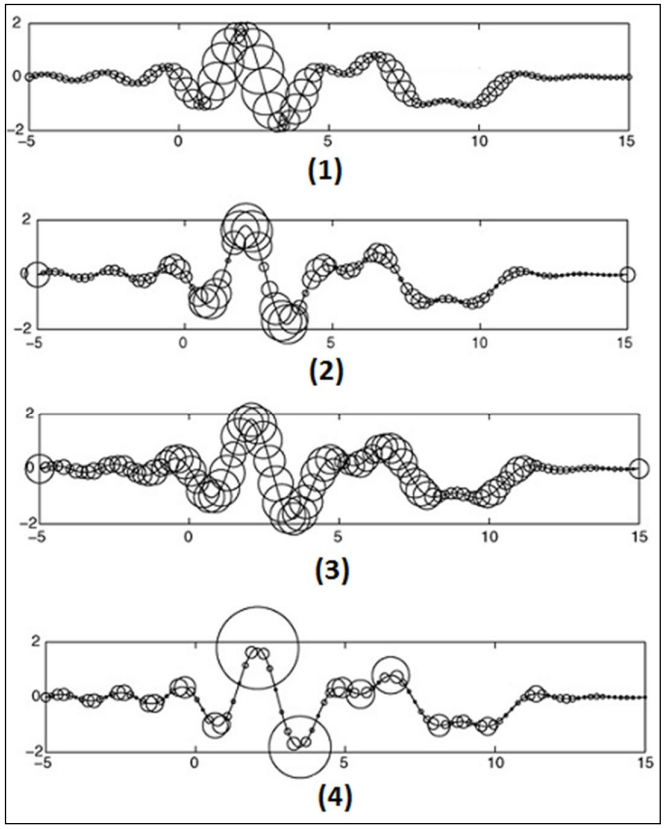
Cada esquema de ponderação apresentado anteriormente pode ser usado para classificar os pontos de dados a partir dos menos relevantes (menor importância), em direção aos mais relevantes (maior importância). A maneira trivial para reduzir o conjunto de dados e

aumentar a qualidade, seria reduzir o conjunto de dados em determinada porcentagem de exemplos, retirando aqueles que possuem menor peso. No entanto, isso não seria a solução mais adequada de comprimir, devido à maneira com que são definidos os pesos.

Os altos valores da proximidade, afastamento e os pesos através do desvio no hiperplano irão detectar os pontos localizados em área esparsa ou em áreas remotas do espaço de dados. Os pontos com muito baixo peso podem estar localizados perto de seus vizinhos mais próximos ou estar em uma relação linear com eles. Assim, um exemplo do conjunto de dados recebe um baixo peso quando ele está localizado em um pequeno aglomerado denso, cujo tamanho é um pouco maior que o tamanho da vizinhança usada na definição de peso.

Na Figura 8 é apresentado um exemplo da utilização dos quatro métodos de construção de pesos em um conjunto de dados sintético. Nela é possível verificar que os métodos 2, 3 e 4 auxiliam com maior eficiência na identificação dos pontos de baixa incidência no conjunto, como os pontos nos índices 2 e 4 da base. O primeiro, o segundo e o terceiro método apresentam uma maior dificuldade em representar os pontos que estão em áreas sobre representadas no espaço.

Figura 8 - Comparação entre os quatro métodos de construção de pesos apresentados.



Fonte: Adaptado de Vladislavleva et. al. (2010).

Para utilizar os esquemas de ponderação apresentados, Vladislavleva et al. (2010) aborda em seu trabalho a aplicação em uma PG, a qual é alterada a fim de melhorar o desempenho das soluções geradas utilizando dados desbalanceados. Para realizar tal tarefa a função de aptidão da PG utiliza o conjunto de pesos para adaptar a aptidão do indivíduo, multiplicando o exemplo avaliado pelo seu respectivo peso durante o processo de treinamento.

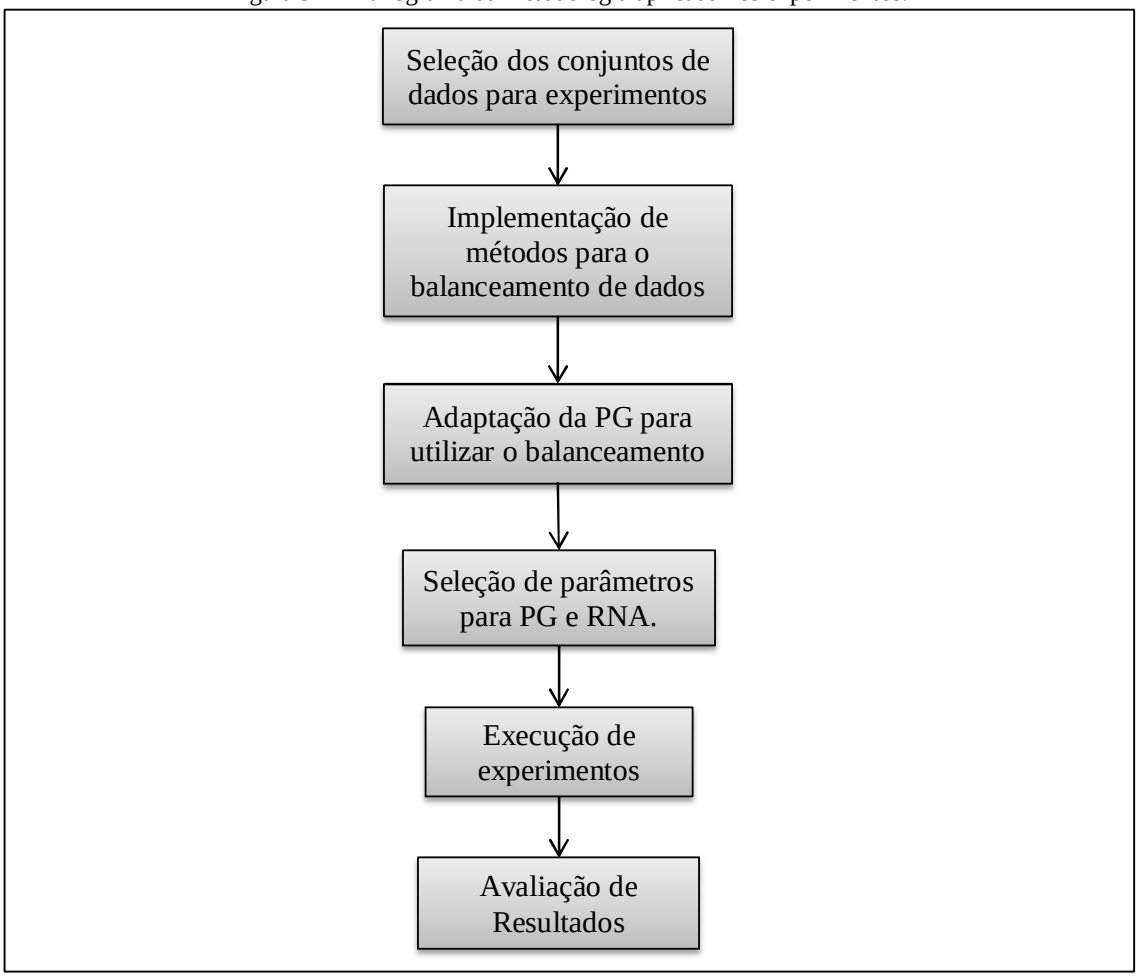
Nesta dissertação os quatros métodos de ponderação foram implementados, junto com a adaptação necessária nas funções de aptidão da PG.

3 EXPERIMENTOS E VALIDAÇÃO

Este capítulo descreve como foram implementadas e utilizadas as técnicas de modelagem PG, RNA e o balanceamento citados nos capítulos anteriores e também descrever como as mesmas foram validadas sob conjuntos de dados experimentais.

Para realizar estes experimentos foi utilizada a metodologia apresentada na Figura 9 em forma de fluxograma.

Figura 9 - Fluxograma da metodologia aplicada nos experimentos.



3.1 CONJUNTOS DE DADOS

Para avaliar a eficiência do método de balanceamento de dados na construção de modelos para PG, foram selecionados seis conjuntos de dados experimentais públicos retirados do repositório *UCI Machine Learning Repository* (FRANK; ASUNCION, 2010). Os conjuntos de dados referem-se a problemas variados que podem ser aplicados em uma tarefa de regressão. Cada conjunto é apresentado a seguir com sua respectiva descrição:

1) Hardware de Computador: Este conjunto de dados refere-se a informações de desempenho de uma CPU (Unidade de Processamento Central, do inglês, *Central Processing Unit*) descritas em termos de tempo de ciclo, tamanho da memória, entre outros. Nos valores de desempenho relativo foram estimados pelos autores da base, utilizando um método de regressão linear (KIBLER ET AL., 1989). Portanto o objetivo de regressão com esta base é prever o desempenho de uma CPU a partir de informações sobre a mesma.

2) Resistência do Concreto à Compressão: Concreto é o material de extrema importância na engenharia civil. A força de compressão do concreto em *MPa* é uma função não linear dada pela idade e ingredientes utilizados no preparo (YEH, 1998). O objetivo da utilização desta base é prever a resistência do concreto em relação à composição e a idade desses materiais.

3) Incêndios Florestais: Esta base é difícil tarefa de regressão, onde o objetivo é prever a área queimada por incêndios florestais, na região nordeste de Portugal, utilizando dados meteorológicos e outras informações (CORTEZ; MORAIS, 2007).

4) Telemonitorização de Parkinson: O conjunto de dados foi criado por pesquisadores da Universidade de Oxford, em colaboração com 10 centros médicos nos Estados Unidos e a empresa *Intel Corporation*, que desenvolveram um dispositivo de telemonitorização para gravar os sinais da fala. O estudo original utiliza uma variedade de métodos de regressão linear e não linear para prever o escore clínico de sintomas da doença de Parkinson na escala UPDRS (LITTLE ET AL., 2007).

5) Qualidade do Vinho: Este conjunto de dados representa informações sobre qualidade do vinho tinto medidas em índices de 0 (ruim) a 10 (excelente). Para prever a qualidade do vinho são utilizados atributos referentes à composição da bebida. (CORTEZ ET AL., 2009).

6) Câncer de Mama Wisconsin (Prognóstico): Cada exemplo desta base representa os dados do acompanhamento de um caso de câncer de mama. Os atributos são calculados a partir de uma imagem digitalizada de uma Punção Aspirativa por Agulha Fina (PAAF) de uma mama (STREET ET AL., 1995).

Os conjuntos de dados apresentados possuem apenas atributos numéricos com objetivo de prever outro atributo numérico, sendo todos configurados para uma tarefa de regressão.

Na Tabela 1 são apresentadas informações sobre cada conjunto de dados como a quantidade de atributos, exemplos, valor mínimo do atributo de predição, valor máximo do atributo de predição, média do atributo de predição e desvio padrão do atributo de predição.

Tabela 1 - Informações sobre os conjuntos de dados experimentais

Base	nº de Atributos	nº de Exemplos	Valor Mínimo	Valor Máximo	Média	Desvio padrão
1	7	209	15	1238	99,330144	154,7571022
2	8	1030	2,33	82,6	35,817961	16,70574196
3	12	517	0	1090,84	12,847292	63,65581847
4	21	5975	0,065937	0,486	0,2105296	0,063408567
5	11	1599	3	8	5,6360225	0,80756944
6	9	699	2	4	2,6895565	0,951272532

Dentre os conjuntos de dados, os conjuntos 1, 2 e 3, destacam-se por possuírem maior característica de dados desbalanceados. Para identificar esta característica eles foram submetidos a um processo de discretização, no qual o atributo alvo foi divido em três classes de intervalos iguais de acordo com o valor máximo e mínimo do atributo. Nestes conjuntos o atributo alvo possui baixas representatividades em determinadas classes, principalmente naquelas que possuem os maiores valores. Por exemplo, no conjunto de dados 1, que possui 209 exemplos, são identificados na primeira classe (15 – 422) 203 exemplos, na segunda (423 - 830) e terceira (831 - 1238) classe apenas 3 exemplos em cada uma.

Os conjuntos 4, 5 e 6 são balanceados, mas com sua devida importância no trabalho, pois foram utilizados na verificação do efeito da aplicação do método de balanceamento sobre dados balanceados.

3.2 METODOLOGIA

Nesta seção descreve-se como as técnicas de PG para dados desbalanceados e RNA foram aplicadas aos conjuntos de dados experimentais.

3.2.1 Implementação

Os criadores do R o definem como uma linguagem e ambiente de programação estatística e gráfica, o R também é um programa “orientado a objeto” (*object oriented programming*. Com o R é possível manipular e analisar dados, visualizar gráficos e escrever

desde pequenas linhas de comando até programas mais completos, como é o caso deste trabalho (R Development Core Team, 2008).

Os quatro métodos de construção de pesos descritos na seção 2.3.1 foram implementados no *software* estatístico R, os quais recebem como parâmetros de entrada o conjunto de dados que se deseja construir os pesos e o tamanho de *k* vizinhos a serem considerados pelos métodos, resultando em um vetor que contem o peso de cada exemplo pertencente ao conjunto de dados.

Os modelos construídos com a PG também foram por meio do *software* R, utilizando um pacote de PG desenvolvido por Flasch et al. (2010), chamado de R *Genetic Programming framework* (RGP). O pacote RGP foi adaptado para que fosse possível alterar a função de aptidão da PG de tal forma, que o conjunto de pesos obtido fosse multiplicado pela aptidão do indivíduo em cada exemplo.

Outra opção implementada para que os experimentos pudessem ser realizados foi o método de validação cruzada. Validação cruzada é um método de validação do modelo, que consiste em dividir o conjunto de dados em x partes, chamadas de partições. Destas, *x*-1 partições são utilizadas para o treinamento, e uma serve como conjunto de testes. O processo é repetido *x* vezes, de forma que cada partição seja utilizada uma vez com o conjunto de testes. Ao final, as medidas de desempenho total são calculadas pela média dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas (FAYYAD, 1996).

A RNA foi utilizada por meio do *software* Matlab (MATHWORKS, 2012) com o *Neural Network Toolbox*. O Matlab é um *software* interativo de alta performance voltado para o cálculo numérico, integrando análise numérica, cálculo com matrizes, processamento de sinais e construção de gráficos em um único ambiente, onde problemas e soluções são expressas de forma matemática, ao contrário da programação tradicional.

3.2.2 Seleção de parâmetros

A fim de definir a melhor configuração para aprendizagem de um modelo utilizando PG, foi realizados testes empíricos com alguns parâmetros de cada técnica. O quadro 1 apresenta os parâmetros que foram ajustados na PG, com os valores que foram testados.

Quadro 1 - Parâmetros de aprendizagem utilizados na PG.

Id	Parâmetro	Valores avaliados
a	Tamanho da população	50, 80, 100, 150, 200, 250 e 300
b	Condição de parada	100, 200, 300, 400, 500
c	Função de aptidão	MSE, MSEW, RMSE, RMSEW, MAE e MAEW
d	Conjunto de funções para construção do indivíduo	1, 2, 3, 4, 5 e 6
e	Taxa de cruzamento	Valores entre 0 e 1
f	Taxa de mutação	Valores entre 0 e 1
g	Método de seleção: tamanho do torneio	Valores entre 2 até o parâmetro "a"

a) Tamanho da população: Tamanho da população de indivíduos a serem evoluídos.

b) Condição de parada: Número de gerações utilizadas como condição de parada na construção do modelo.

c) Função de aptidão: Função utilizada para avaliar o erro da predição do indivíduo.

- 1 – MSE: Erro quadrático médio
- 2 – MSEW: Função adaptada para multiplicar o peso do exemplo pelo MSE.
- 3 – RMSE: Raiz do erro quadrático médio
- 4 – RMSEW: Função adaptada para multiplicar o peso do exemplo pelo RMSE.
- 5 – MAE: Erro médio absoluto
- 6 – MAEW: Função adaptada para multiplicar o peso do exemplo pelo MAE.

d) Conjunto de funções para construção do indivíduo: Funções que são utilizadas para formar o indivíduo. Essas funções são divididas em conjuntos, mas também podem ser concatenadas.

- 1 – Conjunto de funções matemáticas básicas: "+", "-", "*" e "/".
- 2 – Conjunto de funções exponencial e logarítmicas: "*exp*" e "*log*".
- 3 – Conjunto de funções trigonométricas: "*sen*", "*cos*" e "*tag*".
- 4 – Conjuntos de funções matemáticas: os três anteriores juntos.
- 5 – Conjunto concatenado: "+", "-", "*", "/", "*sen*", "*cos*" e "*tag*".
- 6 – Conjunto concatenado: "+", "-", "*", "*sen*", "*cos*" e "*tag*".

e) Taxa de cruzamento: O cruzamento implica na combinação de partes de dois programas para a formação de outros dois indivíduos para a nova população. Esse operador visa criar melhores soluções a partir das existentes, de modo a explorar o espaço de busca de todas as soluções possíveis. Esta taxa pode variar entre 0 e 1, ou seja, entre 0 e 100%.

f) Taxa de mutação: O operador de mutação altera partes internas da estrutura do programa para aumentar a diversidade na população. Deste modo, evita a convergência prematura ou o estacionamento em um ponto do espaço de busca em que se encontram programas subótimos. Esta taxa pode variar entre 0 e 1, ou seja, entre 0 e 100%.

g) Método de seleção: Foi estipulado o torneio como método de seleção, visando controlar a pressão seletiva. Este parâmetro define a quantidade de indivíduos no torneio.

A RNA foi construída utilizando dois algoritmos de aprendizagem diferentes, o *backpropagation* e o *resilient propagation*, com suas particularidades nos parâmetros. Para o *backpropagation* foi adotado uma taxa de aprendizagem de 0,3 e momento de 0,2. Já para o *resilient* foi adotado os parâmetros de delta inicial ($\Delta_0 = 0.1$), delta máximo ($\Delta_{\max} = 50.0$), delta mínimo ($\Delta_{\min} = 1e^{-6}$), fator de redução ($\eta^- = 0.5$) e fator de incremento ($\eta^+ = 1.2$).

Na RNA os modelos foram construídos utilizando apenas uma camada oculta. O número de épocas e o número de neurônios na camada oculta foram testados com diferentes configurações, de 500 à 10000 em um passo de 500 para o número de épocas e de 5 à 30 em um passo de 1 para o número de neurônios na camada oculta.

Cada conjunto de parâmetros para a PG e para RNA foi aplicado apenas no primeiro conjunto de dados, o conjunto de Resistência do Concreto à Compressão, e o melhor desempenho foi replicado para os demais conjuntos. Para a PG o melhor conjunto de parâmetros encontrado foi (a=100, b=200, c=4, d=6, e=0.9, f=0.1, g=6). Para a RNA a melhor configuração foi de 2000 épocas e 15 neurônios ocultos.

3.2.3 Execução dos experimentos

Os experimentos descritos aqui visaram avaliar as diferentes técnicas de PG, RNA e balanceamento para estudos de casos experimentais, para assim identificar o desempenho das mesmas.

Para aplicar a PG com balanceamento nos conjuntos de dados é necessário estabelecer o conjunto de pesos para cada base. Dentre os quatro métodos propostos, segundo Vladislavleva et al. (2010) o que melhor se adapta na maioria dos conjuntos de dados é o terceiro método, pelo afastamento de *k* vizinhos em um espaço de vizinhos, o qual foi adotado nos seis conjuntos de dados, com o valor de *k* igual a 4.

Com o conjunto de pesos definidos foram realizadas quatro abordagens diferentes. Para cada abordagem os seis conjuntos de dados experimentais são aplicados utilizando cinco partições de validação cruzada. As abordagens são as seguintes:

1) Aplicação da PG com a adaptação na função de aptidão para que utilize-se o conjunto de pesos estabelecidos (PGB).
2) Aplicação de uma PG sem a utilização de pesos para o balanceamento (PGC).
3) Aplicação de uma RNA com o algoritmo de aprendizagem *backpropagation* (RNAB).
4) Aplicação de uma RNA com o algoritmo de aprendizagem *resilient* (RNAR).

Estas quatro abordagens permitiram identificar em quais casos o conjunto de pesos melhora o desempenho da PG em comparação com as outras, visando assim, validar se a utilização de pesos implementada reduz o efeito do desbalanceamento de dados, melhorando o desempenho obtido nos modelos.

Tanto a PG quanto a RNA, são inicializadas aleatoriamente, portanto podem chegar a soluções diferentes em cada execução com os mesmos parâmetros. Portanto cada abordagem foi executada 30 vezes, para que seja possível identificar com maior precisão quando uma abordagem foi melhor que outra.

Para auxiliar nesta identificação o teste estatístico de Kruskal-Wallis com 95% de confiança foi aplicado sob duas medidas de desempenho do modelo, a Correlação e a Raiz do Erro Quadrático Médio. Nesse caso, o teste de Kruskal-Wallis possibilita identificar quando uma abordagem possui uma distribuição significativamente diferente de outra, em uma determinada medida de desempenho. Para cada medida o teste foi aplicado com o resultado de cada abordagem em todas as bases, totalizando 180 amostras em cada distribuição.

É importante ressaltar que para cada conjunto de dados uma abordagem pode ser escolhida como a melhor. Porem pode não existir uma abordagem melhor para todos os conjuntos de dados, por este motivo para um determinado conjunto de dados, o ideal é analisar diferentes técnicas para obter-se o melhor modelo.

3.3 RESULTADOS

Nesta seção são apresentados os resultados obtidos com a execução de cada abordagem. Na Tabela 2 estão as médias das medidas de Correlação (Cor.), RMSE e MAE obtidas com cada abordagem, sendo destacadas em negrito os melhores resultados em termos absolutos para cada medida de desempenho apresentada em cada conjunto de dados.

Tabela 2 - Resultados obtidos nos conjuntos de dados experimentais.

Abordagem	Medida	1	2	3	4	5	6
PGB	Cor.	0,94	0,94	**0,36**	**0,94**	0,49	0,48
	RMSE	**14,33**	**6,84**	**52,26**	**0,04**	0,89	0,84
	MAE	7,76	**4,32**	26,81	**0,03**	0,74	0,68
PGC	Cor.	0,8	0,49	0,06	0,76	0,11	0,55
	RMSE	20,75	20,03	64,1	0,07	1,7	0,77
	MAE	10,04	14,11	19,27	0,05	1,47	0,6
RNAB	Cor.	**0,99**	0,9	0,04	0,9	0,41	0,86
	RMSE	16,64	8,02	152,35	0,04	1,02	0,54
	MAE	**6,51**	6,2	40,92	0,03	0,73	0,23
RNAR	Cor.	0,99	**0,99**	0,24	0,89	**0,61**	**0,95**
	RMSE	18,13	19,09	53,73	0,04	**0,64**	**0,29**
	MAE	10,49	10,69	**16,54**	0,03	**0,5**	**0,07**

Observando as médias dos resultados da Tabela 2 é possível verificar que a PG com balanceamento (1) obteve melhores modelos em todos os conjuntos de dados comparada a PG comum (2). Outra abordagem que se destacou foi a RNA com o algoritmo de aprendizagem *resilient* (4) que obteve bons resultados em alguns dos casos, entretanto, somente as médias não são suficientes para realizar tal afirmação.

Os resultados da aplicação do teste de Kruskal-Wallis sobre as distribuições de Correlação e RMSE são apresentados nas Tabela 3 e Tabela 4 respectivamente, nas quais o valor **DIFERENÇA**, representa quando houve diferença estatisticamente significativa entre as distribuições das abordagens.

Tabela 3 – Saída do teste de Kruskal-Wallis para Correlação – bases experimentais.

Correlação	PGB	PGC	RNAB	RNAR
PGB	X	**DIFERENÇA**	N	**DIFERENÇA**
PGC		X	**DIFERENÇA**	**DIFERENÇA**
RNAB			X	N
RNAR				X

Tabela 4 - Saída do teste de Kruskal-Wallis para RMSE – bases experimentais.

RMSE	PGB	PGC	RNAB	RNAR
PGB	X	**DIFERENÇA**	N	N
PGC		X	N	**DIFERENÇA**
RNAB			X	**DIFERENÇA**
RNAR				X

Analisando os resultados obtidos com o teste de Kruskal-Wallis e as médias das medidas de desempenho, é possível afirmar que a PG com balanceamento obteve resultados superiores ao desempenho de uma PG sem o balanceamento, para os conjuntos de dados balanceados e desbalanceados.

Em relação a RNA com *backpropagation* a PG foi melhor em alguns casos e pior em outros, conseguindo uma melhor correlação em relação a todos os conjuntos, porem um RMSE não significativo. A utilização da RNA com o algoritmo de aprendizagem *resilient* obteve melhores resultados que as demais técnicas na maioria dos casos com diferença significativa.

O teste de Kruskal-Wallis também foi aplicado separadamente para cada conjunto de dados, porem os resultados encontrados não diferem das conclusões com a aplicação do teste em todos os conjuntos de dados.

4 ESTUDO DE UM CASO REAL: MOFO BRANCO

Este capítulo tem por objetivo aplicar as técnicas de aprendizagem de máquina na construção de um modelo utilizando dados desbalanceados para previsão de escleródios no estado do Paraná, a partir de características meteorológicas e do local da amostra.

Na seção 4.1 apresenta-se uma síntese sobre a doença de plantas agrícolas mofo branco, com seus aspectos principais. A seção 4.1.1 descreve uma breve introdução sobre a doença. A seção 4.1.2 aborda o ciclo de vida da mesma e a seção 4.2.3 alguns trabalhos que já foram desenvolvidos na modelagem da doença.

As seções 4.2 e 4.3 descrevem respectivamente a metodologia e resultados encontrados na criação deste modelo.

4.1 A DOENÇA MOFO BRANCO

4.1.1 Introdução

O patógeno *Sclerotinia sclerotiurum* (Lib.) de Bary é o agente causador da doença Mofo Branco (MB) que também é conhecida por diferentes denominações, dentre elas: podridão Branca da haste da soja, a esclerotinia, murcha de esclerotinia, podridão da haste (PURDY, 1979).

Sclerotinia sclerotiurum pode causar danos em várias plantas de interesse econômico, são mais de 400 espécies, dentre elas a soja, que podem ser hospedeiras do fungo (BOLAND; HALL, 1994). Este patógeno está disseminado por diversos países em todo o mundo e seus danos se manifestam com maior severidade em regiões com um clima úmido e de alta umidade relativa (PURDY, 1979).

Este fitopatógeno poder permanecer no solo por mais de 10 anos, por meio de estruturas de resistência conhecidas como escleródio ou esclerócio (Figura 10), que são duras em um formato irregular, podendo chegar a vários milímetros de diâmetro e comprimento, a princípio apresentam coloração branca, e posteriormente, tornam-se negros e duros (SAGATA, 2010). O escleródio é formado por uma massa compactada e uma camada externa de pigmentos de melanina que dá resistência à ação de microrganismos e que após sua germinação é o responsável inicial pela infecção da planta cultivada. O tempo de sobrevivência deste fungo pode variar de acordo com o tipo do solo, pH, cultura anterior e condições ambientais (SAGATA, 2010).

A doença no Brasil, tem se tornado importante devido a recentes epidemias ocorridas na cultura da soja, que é responsável por grande parte das exportações brasileiras, principalmente em regiões onde ocorrem condições climáticas amenas na safra de verão (Região Sul, chapadas dos cerrados, acima de 800m de altitude) ou mesmo, em anos de ocorrência de chuvas acima da média (EMBRAPA, 2009).

Figura 10 - Escleródios no solo.

Fonte: Adaptado de MARTINS et al. (2009).

4.1.2 O ciclo da doença na planta

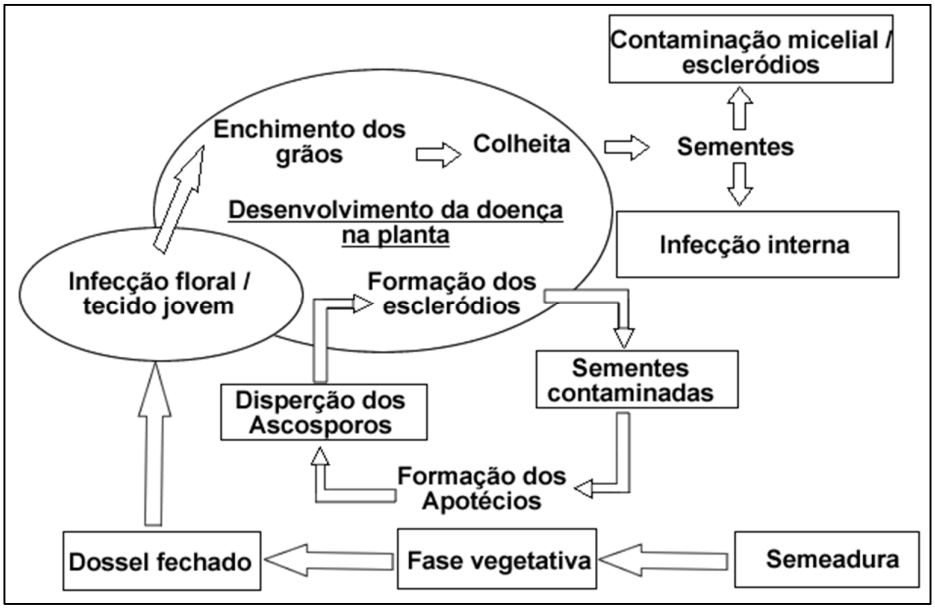
Segundo Juliati e Juliati (2010) que estudaram vários trabalhos da literatura a disseminação do fungo ocorre de duas formas (Figura 11): pela associação do escleródio com as sementes, restos vegetais contaminados na fase miceliogênica do fungo infectando internamente o embrião da semente respondendo pela contaminação a longa distancia, e também se dissemina pelos apotécios que emergem ou brotam dos escleródios do solo, liberando aproximadamente mais de dois milhões de ascósporos por dez a quinze dias, já na fase carpogênica a contaminação ocorre a curtas distancias. A contaminação pode ocorrer também com o transito das maquinas agrícolas que saem das áreas contaminadas para as sadias dentro das lavouras, movimentando o solo e restos vegetais contaminados.

Com a formação dos apotécios a probabilidade de infecção aumenta, e se agrava com o molhamento foliar e a falta de ventilação, ocasionada pelo fechamento precoce da cultura, deste modo favorecendo a ejeção dos ascósporos para infecção das hastes da planta (ABAWI; GROGRAN, 1979).

Pesquisas apontam que a irrigação precoce durante o florescimento aumenta a gravidade da doença nos campos. A fase mais vulnerável da planta durante seu cultivo é no

estádio da floração plena, quando as vagens começam a se formar. Neste momento se as plantas estiverem expostas a altas umidades e temperaturas amenas, o desenvolvimento do fungo é favorecido, caso contrário ele continua no solo até que seja destruído por microrganismos ou encontre a condição ideal (ABAWI; GROGRAN, 1979).

Figura 11 - Ciclo de vida de *Sclerotinia sclerotiorum.*



Fonte: Adaptado do site *White Mold Life Cycle* da Universidade de Minnesota (2011).

Para Costa et. al. (2004), uma das formas eficientes de controlar o MB é por meio da ação de fungicidas, porem a eficiência destes produtos químicos depende de vários fatores, como a densidade de inoculo no solo, o estágio da epidemia, o grau de cobertura das plantas pelo fungicida, o número de pulverizações, a sua fungitoxidade, dose, época, volume, equipamento de aplicação, espaçamento de plantas e severidade da doença. Além de toda esta dificuldade de aplicação e do custo a que ele se associa, também deve-se considerar que qualquer produto químico afetará o solo e o ambiente de alguma forma, por isto a existência de modelos que expliquem a doença ou auxiliem na sua previsão é importante para a obtenção da melhor eficiência em sua aplicação.

Outras formas de controlar a doença é por meio do controle cultural com formação da palhada no sistema de plantio direto associado ao controle biológico, conforme demonstrado em condições experimentais (FERRAZ et al., 1999). Mesmo com resultados positivos obtidos com uso destas práticas não foi comprovada sua eficácia no controle da doença, nem na palhada (controle cultural) e nem no controle biológico. Além desta, existem várias praticas como: utilização de sementes sadias, rotação de culturas (gramíneas), manejo da cultura (espaçamento / população de plantas), aplicação de fungicidas, utilização de

agentes biológicos, eliminação de plantas daninhas, trânsito e limpeza de implementos (JULIATI; JULIATI, 2010).

Mesmo utilizando todas estas práticas os resultados ainda não conseguem garantir e proporcionar benefícios para as safras subsequentes sem repetição dessas estratégias de manejo na mesma área.

4.1.3 Modelagem da doença

A utilização de modelos e da matemática no controle e prevenção de patologias vegetais aumentaram lentamente e os trabalhos neste sentido são recentes, desde o primeiro publicado por Dimond e Horsfall (1965). A seguir é apresentada uma síntese de dois trabalhos que buscam construir modelos para a doença mofo branco.

Harikrishnan e del Río (2008) desenvolveram um modelo de previsão da doença mofo branco na Dakota do Norte para ajudar os produtores na decisão de aplicação de fungicidas. A produção deste modelo foi realizada utilizando análise de regressão logística e inclui as variáveis precipitação total, temperatura média, temperatura mínima e número de dias chuvosos para prever a incidência do mofo branco. O modelo consegue representar 85% da variabilidade existente da doença na cultura. Segundo os autores resultados deste estudo sugerem que estes dados utilizados para prever o risco do mofo branco, podem auxiliar os produtores na tomada de decisões sobre aplicação, ou não de fungicidas para controle da doença.

Clarkson et al. (2007) apresentam um modelo de predição para a produção de apotécios pela germinação de escleródios, que consequentemente irão infectar as plantas com a doença. O modelo baseia-se no pressuposto de que uma fase de condicionamento deve ser completada antes que uma fase de germinação subsequente ocorra. A predição para a produção de apotécios foi bem sucedida, e surge como uma proposta para utilização como parte de um sistema de previsão para o mofo branco na cultura de alface.

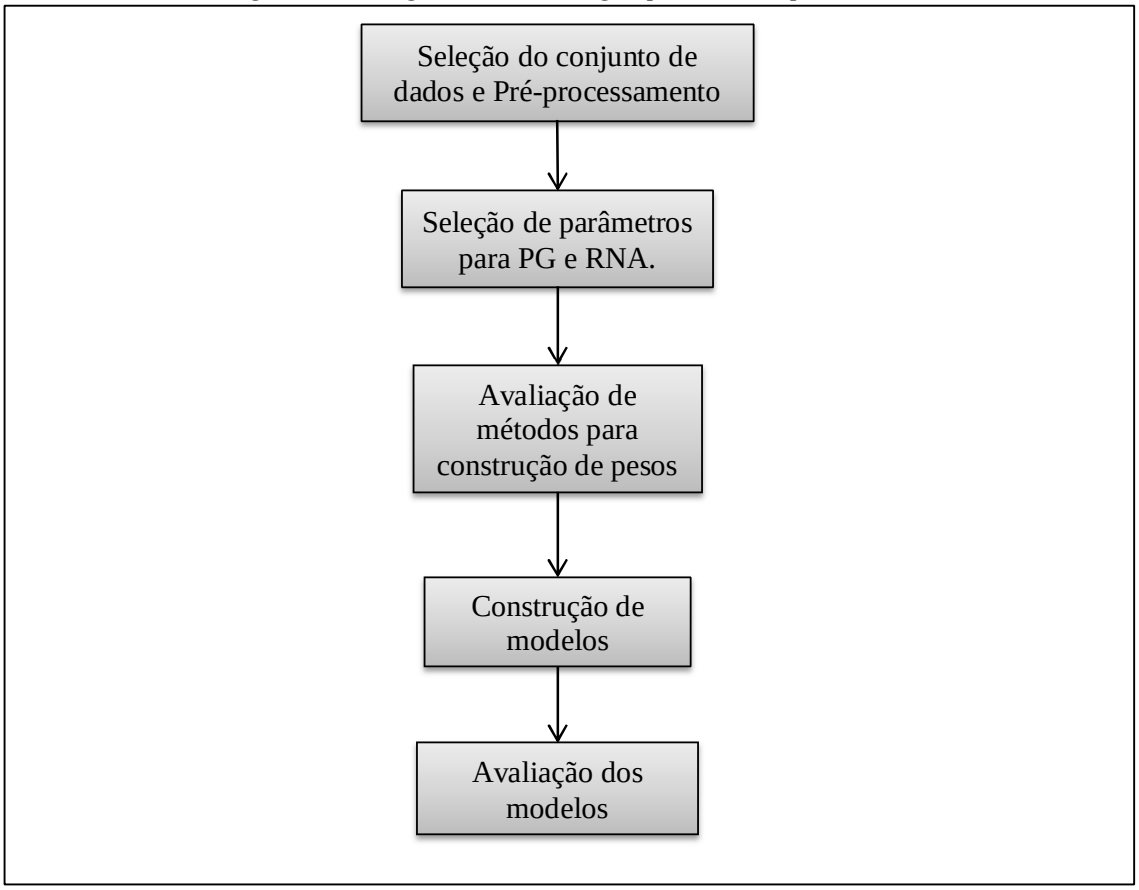
Esta dissertação busca construir um modelo que ainda não foi estabelecido na literatura, que seja capaz de prever a quantidade de escleródios em uma área, de acordo com características meteorologicas e do local da amostra, podendo assim auxiliar na previsão do aparecimento da doença, pois existindo a presença de escleródios no solo a doença pode ocorrer. Com este modelo, será possível prever o aparecimento destas estruturas, auxiliando agricultores na tomada de decisão para o manejo da doença mofo branco.

4.2 METODOLOGIA

Para construir o modelo proposto as duas técnicas de aprendizado de máquina apresentadas nesta dissertação foram utilizadas, a PG e RNA, a fim de encontrar a melhor solução.

Este seção descreve como foram realizados os procedimentos para a criação de um modelo de previsão de escleródios. A metodologia aplicada na construção deste modelo pode ser visualizada na Figura 12 em forma de fluxograma.

Figura 12 - Fluxograma da metodologia aplicada nos experimentos.



4.2.1 Conjunto de dados

O conjunto de dados utilizado neste estudo é parte do Projeto com o título de "Estudos Epidemiologicos e Bio-Moleculares do fungo *Sclerotinia Sclerotiorum* em Sementes e em Áreas de Produção de Soja no Estado do Paraná e Seu Potencial de Risco", aprovado no Edital 064/2008 CNPq/MAPA, com a coordenação do Professor Doutor David de Souza Jaccoud Filho, que junto aos produtores, cooperativas e revendas das regiões estudadas iniciaram uma coleta de informações sobre a doença em locais do estado do Paraná/Brasil, onde já havia ou não ocorrido a doença.

Dentre as informações coletadas, a mais relevante para este trabalho é a quantidade de escleródio presente em uma amostra do solo. Cada amostra de solo de 1 m² foi produzida a partir de quatro coletas de 0,25 m² com profundidade de 5 cm, auxiliada por um molde fixo e uma pá em diferentes lugares de um mesmo talhão. Em seguida estas amostras coletadas forma analisadas em laboratório para se quantificar o número de escleródios existentes no solo. Os dados começaram a ser coletados em 02/12/2008 e até 09/12/2009 o projeto já contava com 480 amostras em 122 áreas, dentro de 67 fazendas distribuídas em 17 municípios (JACCOUD FILHO et al., 2010).

Neste estudo foram utilizadas 179 amostras, onde cada amostra contém informações sobre o local onde foi coletada e o número de escleródio por m² presente. Em cada amostra foram acrescentados os seguintes dados sobre o local: temperatura média mínima de 1960 à 1990, temperatura média máxima de 1960 à 1990, precipitação média de 1960 à 1990 e altitude média do local de cada amostra. Os dados meteorológicos foram obtidos por meio do portal Agritempo (AGRITEMPO, 2011) e a altitude do local através de um GPS (*Global Positioning System*). Uma pequena amostra do conjunto de dados resultante segue na Tabela 5.

Tabela 5 - Amostra do conjunto de dados Mofo Branco

Temp. Min.	*Temp. Máx.*	*Precipitação*	*Altitude*	*n° de Escleródio*
12,3	23,8	1455,3	800	201
12,0	23,6	1382,7	960	56
11,5	23,2	1419,1	985	0
15,8	26,7	1468,8	458	0
15,0	26,8	1667,8	799	150
15,0	26,8	1667,8	787	94
14,8	26,4	1694,2	581	47
13,6	24,1	1470,2	964	88

O objetivo da modelagem neste conjunto de dados foi utilizar a tarefa de regressão para encontrar um modelo que seja capaz de prever o número de escleródios no solo com base nos atributos: temperatura média mínima, temperatura média máxima, precipitação média e altitude.

Este conjunto de dados se caracteriza em um problema de dados desbalanceados. O valor mínimo encontrado para a quantidade de escleródios é de 0 (38 amostras de 179) e o valor máximo de 452 (1 amostra de 179). Para realmente verificar este desbalanceamento, caso o conjunto seja dividido em amostras com no máximo 50 escleródios foi identificado (152 amostras de 179), o que representa a maior parte do conjunto de dados.

4.2.2 Normalização

Uma dificuldade detectada no conjunto de dados abordado é a grande amplitude que existe no atributo n° de escleródio, que varia entre 0 e 452, onde existe poucos pontos com altos números de escleródio e muitos pontos com nenhuma ou pequena quantidade de escleródios. Para tentar amenizar esta amplitude e consequentemente a qualidade do modelo, o atributo escleródio foi normalizado e inserido como mais um atributo no conjunto de dados, Tabela 6.

Tabela 6 - Conjunto de dados mofo branco normalizado.

Temp. Min.	*Temp. Máx.*	*Precipitação*	*Altitude*	*Escleródio Normalizado*
12,3	23,8	1455,3	800	44
12,0	23,6	1382,7	960	12
11,5	23,2	1419,1	985	0
15,8	26,7	1468,8	458	0
15,0	26,8	1667,8	799	33
15,0	26,8	1667,8	787	21
14,8	26,4	1694,2	581	10
13,6	24,1	1470,2	964	19

Esta normalização foi adotada visando reduzir o erro do modelo e com isto dispensar o balanceamento nos dados. Portanto, na tentativa de achar o melhor modelo para o conjunto de dados do mofo branco, este novo conjunto de dados com o número de escleródios normalizado foi adotado como um parâmetro para aprendizagem do modelo.

4.2.3 Parametrização das técnicas

Como apresentado no capítulo anterior a PG com balanceamento obteve melhores resultados que a PG sem o balanceamento nos conjuntos de dados experimentais, e a fim de definir a melhor configuração para aprendizagem de um modelo de previsão de escleródios utilizando PG com balanceamento, parâmetros para criação do modelo foram testados. Além dos parâmetros já apresentados na seção 3.2.2 foram adicionados os parâmetros (h), (*i*) e (*j*), que estão apresentados no Quadro 2.

Quadro 2 - Parâmetros de aprendizagem utilizados na PG.

Id	Parâmetro	Valores avaliados
h	Base utilizada	1 ou 2
i	Método de cálculo dos vizinhos	1, 2, 3, 4
j	Número de vizinhos a serem considerados	1 até o numero de exemplos da base

h) Base utilizada: Utilizando o valor de 1 é adotada base com o número de escleródios real, e 2 para a base com o número de escleródios normalizado.

i) Método de cálculo dos vizinhos: Influencia diretamente no cálculo de pesos, pois ele define como é calculado os pesos para os exemplos do conjunto de dados. Os métodos implementados estão apresentados na seção 2.3.1.

j) Número de vizinhos a serem considerados: Quantidade de exemplos que são considerados como vizinhos para o calculo do peso de um exemplo, este parâmetro é o *k* referenciado na seção 2.3.1.

Novamente a RNA foi construída com dois algoritmos de aprendizagem, o *backpropagation* e o *resilient propagation*, na qual os parâmetros de cada algoritmo foram mantidos conforme a seção 3.2.2. Os modelos foram construídos utilizando apenas uma camada oculta. O número de épocas e o número de neurônios na camada oculta foram testados com diferentes configurações, de 500 à 10000 em um passo de 500 para o número de épocas e de 5 à 30 em um passo de 1 para o número de neurônios na camada oculta.

A definição do conjunto de parâmetros na construção do modelo para previsão de escleródios foi feita por meio de um método de validação, em que o conjunto de dados é fragmentado em dois subconjuntos, denominados, base de treinamento e base de testes. Em uma primeira etapa um algoritmo de aprendizagem é aplicado à base de treinamento. Com isto obtém-se um modelo treinado, que representa o conhecimento extraído dos dados. Na segunda etapa o modelo obtido é aplicado ao fragmento da base de dados denominado base de testes. Como a base de testes também é previamente rotulada, ou seja, possui o valor real do

atributo a ser previsto, é possível medir a taxa de acerto do modelo, comparando a previsão obtida com a previsão real (CABENA et al., 1997). Utilizando este método de validação que é processado em um menor tempo do que na validação cruzada, foram realizados 612360 experimentos para a PG e 120 para a RNA, cada experimento é executado uma única vez com uma diferente combinação de parâmetros, com a intenção de encontrar o melhor conjunto de parâmetros.

Cada modelo foi avaliado sob o resultado de sua predição, considerando as medidas de Correlação e Erro Médio Absoluto (MAE). Na Tabela 7 e Tabela 8 é possível visualizar as 10 melhores configurações encontradas para a PG e RNA respectivamente.

Tabela 7 - Melhores configurações da PG.

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*Correlação*	*MAE*
100	*200*	*4*	*6*	*0,9*	*0,1*	*2*	*1*	*1*	*4*	*0,79*	*33,06*
100	*200*	*2*	*6*	*0,9*	*0,2*	*4*	*2*	*1*	*4*	*0,79*	*26,15*
100	*200*	*4*	*5*	*0,7*	*0,1*	*2*	*2*	*1*	*2*	*0,77*	*17,99*
100	*100*	*2*	*4*	*0,8*	*0,1*	*4*	*1*	*1*	*4*	*0,77*	*17,99*
50	*200*	*4*	*5*	*0,9*	*0,1*	*4*	*1*	*1*	*8*	*0,77*	*46,57*
50	*100*	*4*	*6*	*0,7*	*0,2*	*2*	*2*	*1*	*10*	*0,77*	*30,77*
200	*100*	*2*	*6*	*0,9*	*0,1*	*2*	*1*	*1*	*2*	*0,76*	*53,39*
200	*200*	*2*	*5*	*0,8*	*0,3*	*2*	*1*	*1*	*4*	*0,76*	*50,63*
200	*200*	*4*	*6*	*0,8*	*0,1*	*4*	*1*	*1*	*4*	*0,76*	*45,65*
200	*200*	*2*	*5*	*0,7*	*0,2*	*2*	*1*	*1*	*6*	*0,76*	*54,30*

Tabela 8 - Melhores configurações da RNA.

Algoritmo	*h*	*Número de Neurônios*	*Número de Épocas*	*Correlação*	*MAE*
backpropagation	1	6	2500	0,9164	28,8474
resilient	1	6	2500	0,8237	7,8269
resilient	1	6	1000	0,7576	9,2686
resilient	2	6	1500	0,6834	10,4824
backpropagation	1	10	500	0,4921	8,6467
backpropagation	1	5	500	0,4216	10,2827
resilient	1	5	1000	0,3832	22,2561
backpropagation	1	8	500	0,3036	18,972
backpropagation	2	4	500	0,1869	8,9916
resilient	1	8	500	0,1788	15,8826

Analisando os resultados obtidos com cada configuração foi selecionado o melhor conjunto de parâmetros para PG e RNA. Na PG a melhor configuração encontrada foi (a=100, b=200, c=4, d=6, e=0.9, f=0.1, g=2, h=1, i=2, j=4) e na RNA foi (h=1, Número de épocas igual a 2500 e 6 neurônios ocultos) para o *backpropagation* e *resilient.* Com estes conjuntos

de parâmetros definidos foram aplicadas as mesmas configurações na construção de modelos utilizando o método de validação cruzada.

4.2.4 Criação do modelo

A construção do modelo de previsão de escleródios foi feita por meio da validação cruzada, pois utiliza todo o conjunto de dados para treinar e validar o modelo, enquanto o outro método de validação utilizado para definição dos parâmetros retira amostras da base, que podem ser tendenciosas e levar o modelo a uma previsão muito específica de alguns exemplos.

Para o conjunto de dados do mofo branco o melhor conjunto de pesos identificado na Tabela 7 foi utilizando o método pela proximidade de quatro vizinhos em um espaço de vizinho.

Com o conjunto de pesos definidos, foram aplicadas as quatro abordagens (PGB, PGC, RNAB e RNAR) estabelecidas no capítulo anterior no conjunto de dados do mofo branco, projetadas para serem executas 30 vezes utilizando cinco partições de validação cruzada, e assim identificar com maior precisão qual foi a melhor abordagem aplicada.

O teste estatístico de Kruskal-Wallis foi aplicado sobre a Correlação e a Raiz do Erro Quadrático Médio dos resultados obtidos com as abordagens utilizando 95% de confiança.

4.3 RESULTADOS

4.3.1 Parâmetros das técnicas

Observando os resultados obtidos nos testes que avaliaram os parâmetros da PG (Tabela 7) é possível observar que o parâmetro *h*, responsável por selecionar o conjunto de dados utilizado na aprendizagem, obteve melhores resultados tanto para o conjunto de dados com o atributo escleródio normalizado, quanto ao conjunto em que ele se manteve com o valor real. Isto não ocorreu apenas nos dez melhores modelos, mas sim dentre todos os resultados, ou seja, a normalização do atributo escleródio não resultou em uma melhora significativa sobre os resultados. Portanto, o atributo escleródio normalizado foi descartado da criação do modelo.

A alteração no tamanho da população da PG também não influenciou significativamente nos resultados. Observa-se que nos dez registros apresentados na Tabela 7 o número de indivíduos não influenciou em obter um modelo de qualidade superior.

O conjunto de funções 6 para a construção do indivíduo também é responsável pelos melhores resultados, demonstrando que este não é um simples problema de regressão, e que depende de muitos operadores matemáticos para se chegar a um resultado satisfatório.

As taxas de cruzamento e mutação não influenciaram significativamente na obtenção de melhores resultados, mostraram-se muito variáveis. O padrão que se pode notar é que a taxa de cruzamento que obteve melhores resultados foi com o valor maior ou igual a 0,7 e a taxa de mutação menor ou igual a 0,3. O método para seleção de indivíduos utilizado foi o torneio, que por sua vez também não apresentou diferenças significativas no tamanho do torneio para os melhores resultados.

A função de aptidão da PG tem um papel fundamental nestes experimentos, pois é nesta função que o ajuste na importância do exemplo é realizado, por meio dos pesos criados. As funções de aptidão MSE, RMSE e MAE não implementam o ajuste de pesos, já as funções MSEW, RMSEW e MAEW implementam. Portanto, observando os resultados encontrados na criação de modelos, é nítido que a utilização das funções que implementam o balanceamento influenciam sobre o resultado. Para a criação do modelo de escleródios foi adotado a função RMSEW, mas a função MSEW também apresentou bons resultados, visto que o RMSE é formado com base no MSE.

4.3.2 Método de construção de pesos

Nos testes que avaliaram os parâmetros da PG os quatro métodos de construção de pesos foram testados para o conjunto de dados mofo branco e avaliados por meio dos resultados da predição dos modelos gerados. O método que obtive melhor resultado foi o que utiliza a proximidade de quatro vizinhos em um espaço de vizinhos $\pi(\cdot)$.

Por um lado, Vladislavleva et al. (2010) indica que peso por afastamento $\rho(\cdot)$, é melhor que os métodos guiados pela proximidade $\pi(\cdot)$ e pela distância circundante $\sigma(\cdot)$, mas isso não ocorre neste conjunto de dados. Já em relação ao método de desvio em um hiperplano $\mathcal{V}(\cdot)$ o método $\rho(\cdot)$ realmente apresenta melhores resultados.

4.3.3 Modelo para previsão de escleródios

Nesta seção são apresentados os resultados obtidos com a execução de cada abordagem para o conjunto de dados do mofo branco. As médias das medidas de Correlação, MAE e RMSE dos 30 modelos obtidos com cada abordagem estão apresentadas na Tabela 9.

Tabela 9 - Medidas de desempenho dos modelos para previsão de escleródio.

Abordagem	**Correlação**	**RMSE**	**MAE**
PGB	0,610859979	56,30024	45,29939305
PGC	0,250501270	144,449232	47,45068197
RNAB	0,354965073	83,79022	47,45507463
RNAR	**0,649416667**	**47,62815**	**26,61344333**

Observando as médias das trinta execuções de cada abordagem da Tabela 9 é possível verificar que novamente a PG com balaceamento (1), que utilizou o conjunto de pesos para modelar os dados, foi superior a PG comum (2). O melhor resultado obtido foi aplicando uma RNA com o algoritmo de aprendizagem *resilient* (4).

Aplicando o teste estatístico de Kruskal-Wallis sobre as distribuições de Correlação, RMSE e MAE de cada abordagem chega-se aos resultados apresentados na Tabela 10, Tabela 11 e Tabela 12 respectivamente.

Tabela 10 - Saída do teste de Kruskal-Wallis para Correlação – previsão de escleródios.

Correlação - MB	PGB	PGC	RNAB	RNAR
PGB	X	**DIFERENÇA**	**DIFERENÇA**	**N**
PGC		X	**DIFERENÇA**	**DIFERENÇA**
RNAB			X	**DIFERENÇA**
RNAR				X

Tabela 11 - Saída do teste de Kruskal-Wallis para RMSE – previsão de escleródios.

RMSE – MB	PGB	PGC	RNAB	RNAR
PGB	X	**DIFERENÇA**	**DIFERENÇA**	N
PGC		X	**DIFERENÇA**	**DIFERENÇA**
RNAB			X	**DIFERENÇA**
RNAR				X

Tabela 12 - Saída do teste de Kruskal-Wallis para MAE – previsão de escleródios.

MAE – MB	PGB	PGC	RNAB	RNAR
PGB	X	N	N	**DIFERENÇA**
PGC		X	N	**DIFERENÇA**
RNAB			X	**DIFERENÇA**
RNAR				X

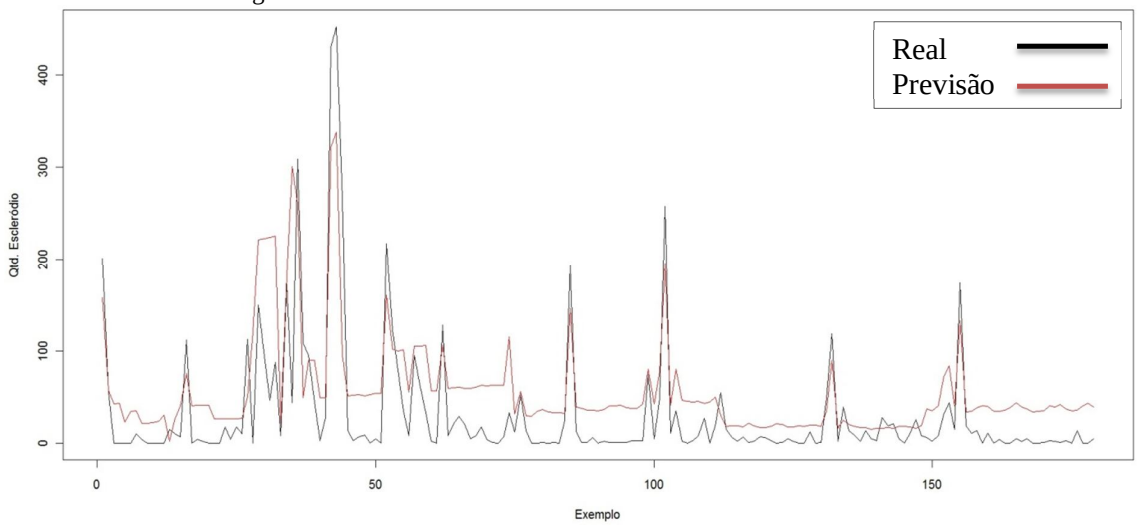
A RNA com o algoritmo *resilient* (4) apresenta um resultado estatisticamente diferente na medida de MAE, na qual as outras abordagens não diferem entre si, pois apresentam as melhores medidas do erro.

Portanto os resultados que chamam a atenção são aqueles obtidos por meio da PG com balanceamento e da RNA com *resilient* e para simular como é o comportamento dos modelos obtidos com estas abordagens foram selecionados dois modelos dentre os 30 gerados para cada abordagem.

Para a PG o melhor modelo obtido possui uma correlação 0,773625, Raiz do Erro Quadrático Médio de 53,22289 e Erro Médio Absoluto 36,47923. A previsão deste modelo pode ser visualizada na Figura 13, que é um indivíduo pertencente a população da PG, representando o comportamento obtido na modelagem dos dados. Este indivíduo é apresentado na equação 17.

Escleródio = Tmax + (cos((sin(Precipitacao + 0.60120)) * (sin(Tmin) + (sin(Precipitacao) - Altitude)) - sin(tan(Altitude) + (Tmax + Altitude) - cos(Tmin) * Precipitacao)) - cos(tan(Altitude))) * tan(sin(Tmin)) * ((tan(Tmin - -0.62261) - sin(Altitude))) * (sin(cos(Precipitacao)) * (sin(Precipitacao) - (Tmin + Tmax))) * (tan(((3.24448 - Tmax) * tan(Tmax) - cos(Precipitacao^2)) * cos(sin(-2.38095 - Tmax))) * cos(sin(Tmax))) * sin((tan(-0.46934) + - 0.95359 * Tmin + cos(sin(Precipitacao))) * (Tmax + 0.008356) + (Tmax + Altitude - sin(Tmax))) * cos(-1.140896* Tmax - cos(Precipitacao)) + (-2.02118 * Precipitacao) (17)

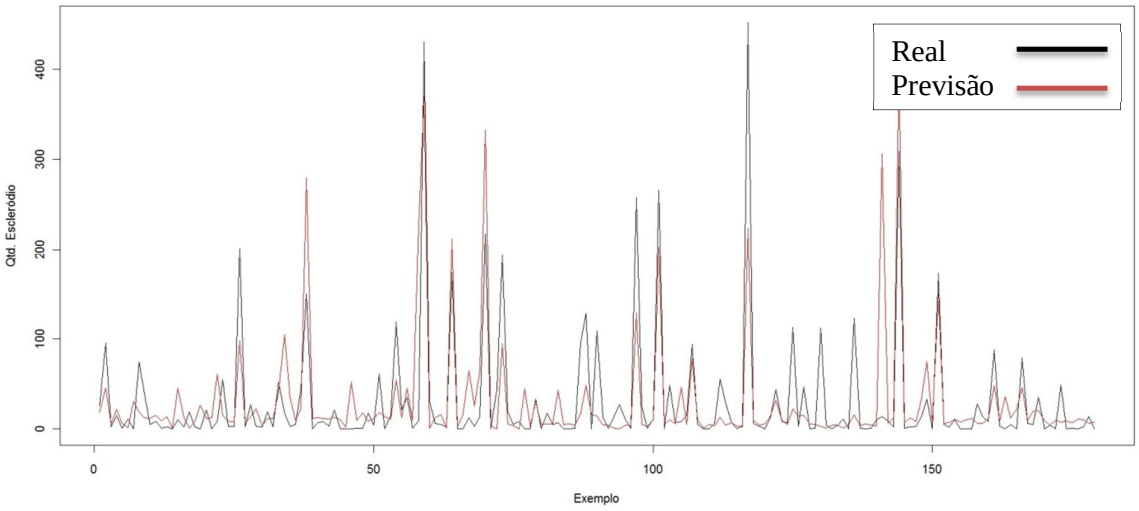
Figura 13 - Previsão do modelo utilizando PG com balanceamento.



A previsão da PG com balanceamento visualizada na Figura 13 representa duas questões importantes a serem levantadas. A primeira é que apesar do conjunto de dados apresentar poucos pontos com alta quantidade de escleródios a PG conseguiu obter um bom índice de acerto nestes pontos, o que é uma grande dificuldade na previsão de dados desbalanceados. A outra questão a ser levantada é que apesar do modelo ter esta melhora, ele não conseguiu prever com qualidade os pontos de baixa ou nenhuma quantidade de escleródios. Entretanto este modelo obtido com a PG utilizando os métodos de balanceamento foi superior a uma PG comum, ressaltando a importância da aplicação de tal técnica.

Este problema na previsão apresentada pela PG não foi encontrado na RNA com o algoritmo de aprendizagem *resilient*, que produziu o melhor modelo para este conjunto de dados. O modelo selecionado dentre os 30 treinados foi o que possui uma correlação 0,763, Raiz do Erro Quadrático Médio de 48,5392 e Erro Médio Absoluto 24,351. Este modelo reduz muito o erro médio da previsão em relação a PG com balanceamento, o que ressalta a qualidade do mesmo. A previsão deste modelo pode ser visualizada na Figura 14.

Figura 14- Previsão do modelo utilizando RNA com dados balanceados.



O resultado do modelo que é representado Figura 14 consegue um melhor acerto médio nos exemplos que possuem um número de escleródios menor que 50 (MAE = 12,7552), e para os exemplos que possuem uma quantidade de escleródios maior ou igual a 50 o erro médio aumenta (MAE = 80,50942).

Os resultados obtidos durantes na criação do modelo para previsão de escleródio induzem a uma avaliação positiva nos modelos encontrados e levantam algumas hipóteses que

poderiam aprimorar o resultado deste modelo. A primeira hipótese a ser levantada é se os dados utilizados para prever a quantidade de escleródios em um local são suficientes, pois a amostra coletada pode não ser representativa o suficiente para o local.

A outra hipótese é que provavelmente existem mais informações que podem auxiliar na criação de um modelo de maior precisão. Abaixo segue algumas variáveis que podem agregar informação aos dados:

1) Umidade Relativa do Solo: Como descrito na seção 4.1 esta é uma variável relevante para o aparecimento da doença, e consequentemente de um número maior de escleródios, pois em condições de alta umidade e temperaturas amenas o desenvolvimento da doença é favorecido. Entretanto esta é uma variável de difícil obtenção.
2) Época de Semeadura: É uma variável que influencia no aparecimento de escleródios do solo, pois conforme a data de semeadura da lavoura pelo agricultor, a planta poderá ser exposta a uma determinada condição climática e consequentemente pelo ciclo da doença poderá gerar mais escleródios. Entretanto esta também é uma variável de difícil coleta, pois muitos agricultores aplicam diferentes datas de semeadura no estado do Paraná/Brasil.

5 CONCLUSÕES

Este trabalho abordou diferentes técnicas de modelagem demonstrando que se deve sempre existir uma preocupação na criação de modelos por meio de técnicas de aprendizagem para dados desbalanceados. Aqui foram destacados os principais aspectos da utilização de dados desbalanceados na criação de modelos, avaliando e comparando duas técnicas de modelagem, PG e a RNA, aplicadas a conjuntos de dados públicos experimentais. Tanto a PG quanto a RNA são técnicas de modelagem muito utilizadas e de resultados já comprovados. A PG foi aplicada neste trabalho utilizando uma abordagem de construção de pesos para reduzir os efeitos do desbalanceamento dos dados durante o processo de aprendizagem, assim resultando em modelo de maior qualidade.

Os testes de validação de PG com balanceamento e comparação das técnicas demonstraram que PG com balanceamento obtêm resultados superiores ao desempenho de uma PG sem o balanceamento.

A utilização da RNA com o algoritmo de aprendizagem *resilient* obteve melhores resultados que as demais abordagens na maioria dos estudos de casos, e em alguns deles com diferença significativa para a medida de MAE.

O estudo de caso que propôs a construção de um modelo de previsão de escleródios obteve bons resultados na utilização de uma RNA com algoritmo de aprendizagem *resilient*, com modelos que chegam a uma correlação de 0,763, Raiz do Erro Quadrático Médio de 48,5392 e Erro Médio Absoluto 24,351. Estes modelos conseguiram prever os exemplos com baixa representatividade no conjunto de dados, quanto os exemplos com alta representatividade, o que é um problema quando se constrói modelos utilizando conjuntos de dados desbalanceados.

5.1 TRABALHOS FUTUROS

Os experimentos realizados nesta dissertação são otimistas, principalmente em relação à modelagem dos dados para previsão de escleródios, tratando-se de um problema real de tal dificuldade. Abaixo são apresentados alguns trabalhos futuros que poderão ser aplicadas nas metodologias estudadas:

- Utilizar outras funções de aptidão na construção do modelo de precisão de escleródios utilizando PG.
- Estudar e implementar outros métodos para balanceamento de dados.

- Visto que a adaptação da PG para utilizar um conjunto de pesos durante o processo de aprendizagem na função de aptidão foi positiva, é possível aplicar este mesmo método a outras técnicas que trabalhem com o aprendizado por minimização do erro de treinamento, como a RNA. Em redes neurais, por exemplo, o conjunto de pesos poderia ser aplicado dentro do algoritmo de aprendizagem da rede, fazendo com que a correção do peso da sinapse fosse realizada por meio da importância do exemplo que foi submetido a rede.
- Agregar em um modelo de previsão de escleródios a variável de umidade relativa do solo, que é de grande importância, porem de difícil obtenção.
- Utilizar as condições meteorológicas de acordo com a época de semeadura da cultura, porem para que isso possa ser realizado, é necessário que haja um histórico de semeadura de cada amostra.

REFERÊNCIAS

ABAWI, G. S. GROGRAN, R. G. **Epidemiology of diseases caused by Sclerotinia species**. Phytopathology, Saint Paul, v. 69, p. 899-910, 1979.

AGGARWAL, C. C.; YU, P. S.**Outlier detection for high-dimensional data**. ACM Special Interest Group Manage. Data Int. Conf. Manage. Data, New York: ACM, 2001, pp. 37–46.

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. **On the surprising behavior of distance metrics in high-dimensional space.** in Proc. Int. Conf. Data Theory, Lecture Notes in Computer Science. 2001, vol. 1973, pp. 420–434.

AGRITEMPO. **Agritempo Site**. url:http://www.agritempo.org.br. Acesso em novembro 2011.

BANZHAF, W. NORDIN, P. KELLER, R. E. FRANCONE, F. D. **Genetic Programming: an introduction**. Morgan Kaufmann, 1998.

BARNETT, V. E; LEWIS, T. **Outliers in Statistical Data**. 3rd Edition. John Wiley and Sons. 1994.

BOLAND, G. J.; HALL, R. **Index of plant hosts of Sclerotinia sclerotiorum**. Canadian Journal of Plant Pathology, Ottawa, v. 16, p. 93 – 108, 1994.

BRAGA, A. DE P.; LUDERMIR, T. B. **Redes Neurais Artificiais Teoria e aplicações**. LTC. Rio de Janeiro. 2000.

CARDIE, C.; NOWE, N. **Improving minority class prediction using case-specific feature weights.** 14th Int. Conf. Mach. Learning, San Francisco, CA: Morgan Kaufmann, pp. 57–65, 1997.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J E ZANASI, A. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1997.

CHERKASSKY, V.; MULIER, F. **Learning from Data: Concepts, Theory, and Methods**. The Wiley bicentennial-knowledge for generations. John Wiley & Sons. 2007.

CHION, C.; LANDRY, J.-A.; DA COSTA, L. **A genetic-programming-based method for hyperspectral data information extraction: Agricultural applications**. Geoscience and Remote Sensing, IEEE Transactions on, 46(8) : 2446–2457. 2008.

CLARKSON, J. P., PHELPS, K., WHIPPS, J. M., YOUNG, C. S., SMITH, J. A., AND WATLING, M. **Forecasting sclerotinia disease on lettuce: A predictive model for carpogenic germination of sclerotinia sclerotiorum sclerotia**. Phytopathology, 97 (5):621, 2007.

CORTEZ, P.; CERDEIRA, A.; ALMEIDA, F.; MATOS, T.; REIS, J. **Modeling wine preferences by data mining from physicochemical properties**. Decision Support Systems, 47(4):547 – 553. Smart Business Networks: Concepts and Empirical Evidence. 2009.

CORTEZ, P.; MORAIS, A. **A data mining approach to predict forest fires using meteorological data**. In 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, pages 512–523. 2007.

COSTA, G.R; COSTA, J. L. da S, **Efeito da Aplicação de Fungicidas no Solo Sobre a Germinação Carpogênica e Miceliogênica de Escleródios de *Sclerotinia Sclerotiorum***. Revista de Pesquisa Agropecuária Tropical, 34 (3): 133-138, 2004.

COSTA, N. C. **Prospecção para a safra 2009/10 – Soja.** Brasília. 2009.

DABHI, V.; VIJ, S. **Empirical modeling using symbolic regression via postfix genetic programming**. In Image Information Processing (ICIIP), 2011 International Conference on, pages 1–6. 2011.

DARWIN, C. **On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life**. Murray, London, UK, 1859.

DIEDERIK, P. LACROIX, R. LEFEBVRE, D. WADE, K. M. **Performance analysis for machine-learning experiments using small data sets.** Computers and Electronics in Agriculture, v. 38, 2003. 1-17.

DIMOND, A. E.; HORSFALL, L. G. **The theory of inoculum**. In Ecology of Soilborne Plant Pathogens, pp. 404- 15, 1965.

EL-TELBANY, M; WARDA, M; EL-BORAHY, M. **Mining the classification Rules for Egyptian Rice Disease**. The International Arab Journal of Information Technology, vol. 3, n. 4, 2006.

EMBRAPA SOJA. **Tecnologias de Produção de Soja: região central do Brasil 2009 e 2010**. Londrina, 2009 Disponível em: <http://www.cnpso.embrapa.br/download/Tecnol2009.pdf.> Acesso em 09 jul. 2011.

FAYYAD, U.M.; PIATETSKI-SHAPIRO, G.; SMYTH, P; Uthurusamy, R. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, p. 11-34. 1996.

FERRAZ, L.C.L.; CAFÉ FILHO, A.C.; NASSER, L.C.B.; AZEVEDO, J.A. **Effects of soil moisture, organic matter and grass mulching on the carpogenic germination of sclerotia and infection of bean by Sclerotinia sclerotiorum**. Plant Pathology, v.48, p.77-82, 1999.

FERNEDA, E. **Redes neurais e sua aplicação em sistemas de recuperação de informação**. Ciência da Informação, Vol. 35, No 1. Brasília. 2006.

FLASCH, O., BARTZ-BEIELSTEIN, T., KOCH, P., AND KONEN, W. **Clustering based niching for genetic programming in the R environment**. In Hoffmann, F. and H¨ ullermeier, E., editors, Proceedings 20. Workshop Computational Intelligence, p. 33–46. Universitatsverlag Karlsruhe. 2010.

FRANK, A.; ASUNCION, A. CI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2010.

GARCIA, R**. Produção de inóculo, efeito de extratos vegetais e de fungicidas e reação de genótipos de soja à Sclerotinia sclerotiorum**. 154 p. Dissertação (mestrado) – Universidade Federal de Uberlândia, Uberlândia, 2008.

GOLDBERG. D. E. **Genetic Algorithms in Search Optimization and Machine Learning**. Addison-Wesley, 1989.

HARMELING, S. DORNHEGE, G. TAX, D. MEINECKE, F. MULLER, K. R. **From outliers to prototypes: Ordering data**. Neurocomputing, vol. 69, nos. 13–15, pp. 1608–1618, Aug. 2006.

HAYKIN, S. **Redes Neurais Princípios e Prática. Bookmark**. Porto Alegre : Bookman. 2001.

HARIKRISHNAN, R.; DEL RÍO, L. E. **A Logistic Regression Model for Predicting Risk of White Mold Incidence on Dry Bean in North Dakota**. Plant Disease, vol. 92, n. 1, 2008. p. 42-46.

HOLLAND J. H., **Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence**. Ann Arbor, MI: University of Michigan Press. 1975.

JACCOUD FILHO, D. S.; MANOSSO NETO, M.; VRISMAN, C. M.; HENNEBERG, L.; GRABICOSKI, E. M. G.; PIERRE, M. L. C.; BERGER NETO, A.; SARTORI, F. F.; DEMARCH, V. B.; ROCHA, C. H. **Análise, Distribuição e Quantificação do Mofo Brando em Diferentes Regiões Produtoras do Estado do Paraná**. In: XXXI Reunião De Pesquisa de Soja da Região Central do Brasil, 2010, Brasília : Embrapa Soja, 2010. p. 226-228.

JI, B.; SUN, Y.; YANG, S.; WAN, J. **Artificial neural networks for rice yield prediction in mountainous regions**. The Journal of Agricultural Science, 145. p. 249–261. 2007.

JULIATTI, F. C.; JULIATTI, F. C. **Podridão Branca da haste da soja: Manejo e uso de fungicidas em busca da sustentabilidade nos sistemas de produção**. Uberlândia, MG, Composer Gráfica e editora Ltda. 33 p. 2010.

KIBLER, D.; AHA, D. W.; ALBERT, M. K. **Instance-based prediction of real-valued attributes**. Comput. Intell., 5(2):51–57. 1989.

KOZA, J. R. Hierarquical **genetic algorithms operating on populations of computer programs**. Proceedings of the 11th International Joint Conference on Artificial Intelligent. Detroit, MI. Morgan Kaufmann, 1989.

KOZA, J. R., **Genetic Programming: On the Programming of Computers by Means of Natural Selection**. WIT Press, 1992.

KRAMER, M. D.; ZHANG, D. **A Genetic Programming System**. In: The 24th Annual International Computer Software and Applications Conference, p. 614-619, IEEE Press, 2000.

LITTLE, M. A.; MCSHARRY, P. E.; ROBERTS, S. J.; COSTELLO, D. A.; M, I.; COSTELLO, D. A.; MOROZ, I. M. **Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection**. 2007.

MATHWORKS. **MATLAB - The Language of Technical Computing**. Disponível em: http://www.mathworks.com/products/matlab. Acesso em 3 de janeiro de 2012.

MARTINS, M. C.; TAMAI, M. A.; LOPES, P. V. L. **Mofo Branco No Brasil**. IV Simpósio da Cultura da Soja. ESALQ / US. 2009.

MCCULLOCH, W. S.; PITTS, W. H. **A logical calculus of the ideas immanent in nervous activity**. Bulletin of Mathematical Biophysics, n. 5. 1943.

MEIRA, C. A. A. **Processo de Descoberta de Conhecimento em Bases de Dados para a Análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro**. Tese de Doutorado, Campinas-SP, UNICAMP, 2008.

MILA, A. L.; CARRIQUIRY, A. L.; YANG, X. B. **Logistic regression modeling of prevalence of soybean Sclerotinia stem rot in the north-central region of the United States**. Phytopathology, v. 94, 2004, p. 102-110.

MINSKY, M. L.; PAPPERT, S. **Perceptron: an introduction to computational geometry**. Cambridge: MIT Press. 1969.

MUCHERINO, A. et al. **A survey of data mining techniques applied to agriculture**. Artigo, Springer, 2009.

MUELLER, A. **Uma aplicação de redes neurais artificiais na previsão do mercado acionário**. Dissertação de Mestrado. Universidade Federal de Santa Catarina. 1996.

MOTULSKY, H. J.; BROWN, R. E. **Detecting outliers when fitting data with nonlinear regression: A new method based on robust nonlinear regression and the false discovery ra**te. BioMed Central Bioinformatics, vol. 7, p. 123, Dez. 2006.

NUNES JR., J. **Relatos por Estado sobre o Comportamento da Cultura de Soja na Safra 2007/2008 : Goiás**. XXX Reunião de Pesquisa da Soja da Região Central do Brasil, 2009. Londrina, Empraba-Soja, 2009. Sessão 2.4, p. 27.

PERINI, A., SUSI, A., **Developing a Decision Support System for Integrated Production in Agriculture.** Environmental Modelling & Software 19, 821-829, 2003.

PEARCE, J. **Programming and Meta-Programming in Scheme**. Birkhäuser, 341 p. 1998.

POLI , R., LANGDON, W. B., MCPHEE, N. F. **A field guide to genetic programming**. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008.

PROVOST, F. J.; FAWCETT, T. **Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions**. 3rd Int. Conf. Knowl. Discovery Data Mining, Newport Beach, CA, 1997, pp. 43–48.

PURDY, L. H. **Sclerotinia sclerotiorum: history, disease and symptomatology, host ranges, geographic distribution and impact**. Phytopathology, Saint Pau, v. 69, n. 8, p. 875-880, 1979.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org. 2008.

RIEDMILLER, M; BRAUN, H. **A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm**. Proceedings of the IEEE, Int. conf. Ond Neural Network (ICN). USA. San Francisco. p. 586-591. 1993

ROSENBLATT, F. **The perceptron: a probabilistic model for information storage and retrieval in the brain**. Psychological Review, v. 65. 1958.

SAGATA, E. **Métodos de inoculação e avaliação da resistência de genótipos de soja à *Sclerotinia Sclerotiorum***. Dissertação de Mestrado, Universidade Federal de Uberlândia, MG, Brasil, 2010.

SAGE, A. P. **Concise Encyclopedia of Information Processing in Systems and Organizations**. Pergamon. New York. 1990.

SILVA, L.H.C.P; CAMPOS, H.D. **Eficacia agronômica de fungicidas no controle do mofo branco (*Sclerotinia sclerotiorum)* da soja,** Cv. BRS Valiosa RR. Rio Verde, FESURV, 15 pag. 2007.

STREET, N.; MANGASARIAN, O. L.; WOLBERG, W. H. **An inductive learning approach to prognostic prediction**. In Proceedings of the Twelfth International Conference on Machine Learning, pages 522–530. Morgan Kaufmann. 1995.

VAPNIK, V. N. **An overview of statistical learning theory**. IEEE Trans. Neural Netw. 1999, 10, 988-999.

VLADISLAVLEVA, E., SMITS, G., AND DEN HERTOG, D. **On the importance of data balancing for symbolic regression**. Evolutionary Computation, IEEE Transactions on, 14(2):252 –277. 2010.

YEH, I.-C. **Modeling of strength of high-performance concrete using artificial neural networks**. Cement and Concrete Research, 28 (12):1797 – 1808. 1998.

WEISS, S.M.; KULIKOWSKI, C.A. **Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems**. Morgan Kaufmann San Mateo, CA, 1991.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. Morgan Kaufmann Publishers Inc. San Francisco. CA. 2005.

ZANETTI, A. L. **Relatos por Estado sobre o Comportamento da Cultura de Soja na Safra 2007/2008 : Minas Gerais**. XXX Reunião de Pesquisa da Soja da Região Central do Brasil, 2009. Londrina, Empraba-Soja, 2009. Sessão 2.3, p. 24.

APÊNDICE A – Resultados dos trinta modelos em cada abordagem

Neste apêndice são apresentadas as tabelas com as medidas de desempenho das trinta execuções de cada abordagem em cada conjunto de dados. Os valores apresentados nestas tabelas foram arredondados para fins de apresentação.

Tabela 13 - Resultados obtidos no conjunto de dados de hardware de computador.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,963	14,26	7,79	0,200	17,09	8,31	0,996	18,75	6,97	0,992	20,23	11,00
0,943	14,04	7,72	0,850	21,07	10,18	0,998	16,31	6,17	0,996	14,43	9,20
0,965	14,05	7,63	0,850	21,07	10,18	0,997	18,12	6,45	0,993	16,20	9,82
0,920	14,53	8,06	0,866	21,13	10,30	0,996	19,36	6,93	0,998	12,69	6,70
0,813	14,18	7,84	0,850	21,07	10,18	0,997	15,16	5,47	0,996	16,18	10,17
0,976	14,27	7,77	0,839	21,07	10,24	0,997	16,00	6,06	0,986	34,97	14,84
0,947	14,05	7,67	0,846	21,10	10,26	0,998	17,06	7,42	0,993	16,15	9,49
0,906	14,58	7,83	0,851	21,07	10,17	0,997	16,74	6,39	0,991	20,07	14,51
0,973	14,05	7,63	0,850	21,07	10,18	0,998	11,56	5,49	0,995	16,18	10,25
0,981	14,10	7,69	0,849	21,07	10,18	0,998	12,92	5,60	0,999	7,06	5,20
0,984	14,28	7,79	0,869	21,06	10,23	0,997	18,35	7,01	0,992	19,01	10,50
0,973	14,05	7,63	0,418	20,73	10,20	0,998	20,00	8,05	0,993	17,87	12,44
0,961	14,04	7,61	0,839	21,06	10,28	0,997	18,07	8,69	0,988	23,48	13,33
0,973	14,05	7,63	0,850	21,07	10,18	0,997	14,53	4,75	0,996	14,10	8,34
0,970	14,05	7,63	0,827	20,98	10,15	0,998	13,09	4,95	0,992	13,29	9,44
0,999	14,32	7,87	0,850	21,07	10,18	0,997	19,70	7,80	0,955	30,63	14,57
0,482	19,43	8,69	0,850	21,03	10,24	0,996	19,79	6,17	0,997	12,87	7,89
0,973	14,05	7,63	0,864	20,94	10,19	0,997	19,92	7,40	0,979	28,03	15,99
0,974	14,05	7,58	0,850	21,07	10,18	0,996	16,19	5,91	0,993	18,78	12,23
0,977	14,08	7,68	0,837	21,05	10,23	0,998	15,74	6,85	0,996	13,41	8,40
0,956	14,05	7,64	0,850	21,07	10,18	0,998	15,01	5,44	0,990	19,88	11,08
0,971	14,05	7,63	0,838	21,01	10,19	0,998	13,21	6,85	0,988	22,29	12,62
0,899	14,05	7,66	0,850	21,07	10,18	0,996	18,16	5,87	0,974	36,15	15,46
0,986	14,09	7,70	0,850	21,07	10,18	0,998	19,91	7,42	0,991	22,14	10,62
0,973	14,05	7,63	0,865	21,09	10,20	0,998	17,91	7,88	0,997	11,11	6,53
0,898	14,33	8,26	0,397	16,00	7,29	0,998	14,14	5,88	0,996	11,05	7,38
0,973	14,05	7,63	0,850	21,07	10,18	0,997	15,88	6,46	0,997	10,21	6,78
0,945	14,10	7,67	0,850	21,07	10,18	0,996	15,31	5,84	0,997	11,08	6,41
0,906	14,58	7,83	0,810	21,26	10,34	0,997	15,00	5,67	0,991	17,00	11,88
0,973	14,06	7,65	0,843	20,91	10,07	0,998	17,39	7,34	0,992	17,44	11,55

Tabela 14 - Resultados obtidos no conjunto de dados de resistência do concreto.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,938	9,06	12,03	0,408	16,26	11,12	0,891	8,04	5,86	0,994	15,19	8,78
0,917	4,28	2,57	0,203	21,81	14,98	0,904	7,47	5,78	0,996	15,02	8,72
0,912	8,60	5,16	0,576	19,24	13,47	0,896	8,04	5,94	0,994	13,61	9,99
0,971	6,91	4,15	0,580	22,88	16,15	0,897	8,23	6,32	0,996	18,84	8,69
0,994	6,27	3,76	0,400	24,07	17,90	0,904	8,32	6,60	0,982	22,39	10,48
0,909	6,64	3,99	0,367	23,90	17,67	0,905	7,81	6,23	0,986	27,91	13,56
0,974	4,89	2,94	0,334	17,54	12,01	0,899	8,17	6,37	0,994	17,85	10,31
0,927	8,08	4,85	0,400	24,07	17,90	0,910	7,27	5,60	0,998	10,47	7,32
0,967	5,71	3,43	0,836	19,07	13,25	0,913	7,25	5,60	0,981	29,24	19,60
0,973	4,16	2,49	0,377	16,63	11,45	0,904	9,46	7,83	0,976	28,63	12,32
0,907	6,88	4,13	0,560	15,20	10,12	0,911	7,54	5,86	0,996	10,69	8,02
0,995	7,08	4,25	0,548	19,37	13,60	0,896	8,34	6,20	0,995	16,70	11,49
0,970	6,85	4,11	0,525	19,92	13,62	0,910	8,34	6,66	0,994	14,22	9,31
0,630	6,16	3,70	0,566	17,92	12,32	0,907	7,36	5,74	0,989	21,34	13,40
0,954	7,27	4,36	0,400	24,07	17,90	0,908	7,54	5,87	0,992	17,96	9,92
0,982	5,03	3,02	0,616	22,88	16,34	0,912	8,18	6,51	0,997	11,62	7,61
0,901	7,97	4,78	0,400	24,07	17,90	0,912	7,86	6,13	0,978	26,92	12,36
0,976	6,81	4,09	0,550	18,45	12,76	0,919	7,04	5,58	0,997	11,17	6,31
0,935	6,82	4,09	0,675	20,41	14,47	0,887	8,41	6,21	0,996	18,15	10,12
0,983	5,60	3,36	0,600	18,60	13,03	0,904	8,56	6,95	0,996	18,48	9,30
0,912	7,14	4,28	0,125	19,71	13,72	0,900	7,57	5,77	0,996	14,52	8,71
0,880	7,63	4,58	0,549	20,27	13,93	0,915	7,59	6,00	0,993	17,61	8,47
0,987	6,95	4,17	0,484	18,43	12,81	0,890	7,97	5,89	0,988	21,75	12,15
0,910	5,71	3,43	0,249	19,88	13,60	0,889	8,28	5,87	0,996	12,99	8,98
0,988	8,60	5,16	0,831	18,62	13,01	0,913	7,45	5,88	0,988	32,28	16,52
0,912	8,60	5,16	0,308	18,87	13,10	0,895	9,50	7,46	0,988	29,58	14,41
0,948	6,75	4,05	0,452	18,44	12,80	0,885	8,92	6,60	0,991	23,15	11,83
0,931	8,00	4,80	0,409	23,22	16,54	0,897	8,18	6,12	0,990	21,47	11,78
0,906	7,77	4,66	0,521	19,97	13,96	0,906	8,56	6,73	0,990	14,40	9,79
0,980	6,86	4,12	0,846	17,26	11,77	0,914	7,34	5,86	0,993	18,50	10,53

Tabela 15 - Resultados obtidos no conjunto de dados de incêndios florestais.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,487	46,47	26,00	0,106	64,39	20,65	0,033	176,88	38,63	0,290	53,20	16,96
0,653	48,00	26,55	0,072	63,53	20,61	0,023	101,38	30,70	0,149	45,76	12,92
0,550	49,70	27,54	0,069	64,15	19,35	0,001	162,60	45,01	0,288	63,13	19,11
0,636	55,25	29,76	0,093	64,65	20,37	0,077	172,46	42,36	0,382	50,41	14,93
0,480	39,91	18,58	0,041	64,06	20,17	0,014	158,99	45,19	0,181	52,43	16,02
0,656	65,93	30,81	0,050	64,16	19,26	0,023	204,80	50,99	0,157	47,96	15,69
0,397	52,02	27,65	0,113	64,42	19,65	0,156	158,95	43,35	0,330	51,91	14,71
0,269	64,09	28,99	0,020	63,91	15,25	0,003	108,85	30,62	0,324	61,71	19,17
0,140	55,76	32,94	0,117	64,38	17,92	0,015	244,16	58,60	0,342	53,73	15,00
0,451	51,58	27,91	0,090	64,21	19,48	0,093	142,70	38,59	0,252	46,33	14,10
0,509	45,93	24,24	0,048	63,71	21,54	0,027	190,64	43,34	0,274	51,87	15,80
0,005	50,31	30,55	0,011	63,92	16,39	0,006	88,07	29,37	0,168	51,12	15,89
0,005	50,31	30,55	0,048	63,72	21,49	0,006	219,39	54,20	0,223	53,41	15,36
0,194	48,74	23,60	0,044	63,57	17,68	0,127	187,06	44,56	0,309	53,09	13,36
0,039	63,80	29,23	0,070	64,20	19,48	0,023	146,65	43,70	0,236	49,13	17,03
0,919	60,55	25,53	0,083	64,14	16,69	0,008	146,08	37,41	0,316	52,64	16,10
0,337	59,45	31,70	0,020	63,91	15,25	0,030	160,52	43,68	0,164	52,68	16,27
0,238	51,36	26,30	0,062	63,60	21,12	0,008	197,65	48,52	0,185	52,14	14,96
0,547	49,68	27,53	0,095	64,71	21,23	0,010	116,00	32,40	0,100	71,60	16,40
0,001	47,50	25,20	0,052	64,03	18,95	0,095	114,19	33,78	0,221	49,73	16,57
0,853	48,10	22,75	0,031	63,96	19,68	0,103	138,24	38,45	0,174	61,46	17,03
0,103	32,13	8,23	0,072	63,53	20,61	0,004	127,00	38,93	0,290	53,43	16,84
0,223	60,24	34,16	0,054	63,95	19,69	0,031	128,10	35,26	0,168	56,46	20,84
0,480	45,77	24,65	0,119	64,43	18,09	0,014	113,10	37,79	0,320	52,61	14,93
0,131	51,21	24,07	0,094	64,67	21,66	0,004	169,57	41,82	0,144	53,73	16,53
0,365	66,19	33,55	0,092	65,12	20,96	0,100	130,69	39,61	0,215	59,89	21,01
0,181	49,06	25,07	0,016	63,93	19,25	0,033	123,65	38,21	0,211	50,49	14,08
0,223	43,09	20,65	0,032	64,16	20,42	0,005	97,12	36,02	0,215	53,80	19,78
0,448	55,10	31,13	0,030	64,10	20,06	0,101	165,85	41,26	0,171	48,86	16,15
0,212	60,64	28,74	0,020	63,91	15,25	0,038	179,11	45,28	0,298	57,20	22,61

Tabela 16 - Resultados obtidos no conjunto de dados de telemonitorização.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,971	0,04	0,03	0,816	0,06	0,05	0,890	0,05	0,03	0,894	0,04	0,03
0,960	0,04	0,02	0,800	0,06	0,05	0,918	0,04	0,03	0,886	0,04	0,03
0,995	0,05	0,03	0,952	0,06	0,04	0,902	0,04	0,03	0,888	0,04	0,03
0,815	0,05	0,03	0,771	0,07	0,05	0,909	0,04	0,03	0,875	0,04	0,03
0,986	0,04	0,03	0,695	0,07	0,05	0,914	0,05	0,04	0,910	0,04	0,03
0,958	0,04	0,03	0,959	0,06	0,04	0,891	0,04	0,03	0,893	0,04	0,03
0,984	0,03	0,03	0,749	0,05	0,04	0,910	0,04	0,03	0,901	0,04	0,03
0,975	0,04	0,03	0,772	0,07	0,05	0,912	0,04	0,03	0,889	0,04	0,03
0,822	0,08	0,06	0,932	0,06	0,05	0,900	0,05	0,03	0,904	0,04	0,03
0,997	0,03	0,02	0,939	0,06	0,05	0,919	0,04	0,03	0,885	0,04	0,03
0,996	0,04	0,03	0,961	0,06	0,04	0,912	0,04	0,03	0,899	0,04	0,03
0,988	0,04	0,03	0,852	0,05	0,04	0,912	0,04	0,03	0,889	0,04	0,03
0,905	0,04	0,03	0,919	0,05	0,03	0,890	0,05	0,03	0,885	0,04	0,03
0,901	0,03	0,02	0,695	0,07	0,05	0,912	0,04	0,03	0,884	0,04	0,03
0,883	0,04	0,03	0,945	0,08	0,06	0,914	0,05	0,03	0,898	0,04	0,03
0,986	0,03	0,02	0,030	0,12	0,09	0,886	0,05	0,03	0,895	0,04	0,03
0,919	0,04	0,03	0,558	0,08	0,05	0,893	0,05	0,03	0,858	0,05	0,04
0,890	0,05	0,05	0,837	0,07	0,05	0,903	0,04	0,03	0,885	0,04	0,03
0,974	0,03	0,03	0,933	0,08	0,05	0,904	0,04	0,03	0,903	0,04	0,03
0,967	0,03	0,03	0,903	0,05	0,04	0,898	0,05	0,03	0,891	0,04	0,03
0,837	0,03	0,02	0,200	0,11	0,09	0,908	0,04	0,03	0,888	0,04	0,03
0,940	0,03	0,02	0,667	0,09	0,06	0,914	0,04	0,03	0,894	0,04	0,03
0,939	0,04	0,03	0,772	0,05	0,04	0,874	0,05	0,03	0,888	0,04	0,03
0,845	0,04	0,03	0,777	0,05	0,04	0,897	0,04	0,03	0,900	0,04	0,03
0,995	0,04	0,03	0,035	0,14	0,10	0,888	0,05	0,03	0,905	0,04	0,03
0,987	0,03	0,02	0,978	0,06	0,04	0,897	0,04	0,03	0,904	0,04	0,03
0,820	0,05	0,03	0,965	0,06	0,04	0,905	0,04	0,03	0,887	0,04	0,03
0,957	0,05	0,04	0,786	0,05	0,04	0,912	0,05	0,04	0,900	0,04	0,03
0,963	0,05	0,03	0,846	0,07	0,05	0,901	0,04	0,03	0,892	0,04	0,03
0,962	0,03	0,02	0,875	0,06	0,05	0,909	0,04	0,03	0,883	0,04	0,03

Tabela 17 - Resultados obtidos no conjunto de dados de qualidade do vinho.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,312	0,58	0,49	0,123	1,00	0,79	0,424	0,95	0,68	0,596	0,65	0,51
0,268	1,11	0,75	0,098	2,47	2,33	0,392	1,01	0,71	0,604	0,64	0,50
0,551	0,66	0,52	0,098	2,47	2,33	0,411	0,95	0,68	0,599	0,65	0,51
0,394	0,92	0,79	0,098	1,32	1,13	0,421	0,99	0,71	0,610	0,64	0,50
0,574	0,80	0,66	0,098	2,47	2,33	0,412	1,02	0,72	0,602	0,65	0,52
0,937	0,73	0,59	0,098	2,47	2,33	0,406	1,07	0,78	0,641	0,62	0,48
0,423	0,64	0,51	0,098	2,47	2,33	0,379	0,97	0,72	0,619	0,64	0,50
0,481	0,66	0,53	0,295	0,89	0,74	0,355	1,03	0,73	0,620	0,63	0,50
0,271	0,65	0,51	0,104	1,01	0,80	0,450	0,96	0,68	0,586	0,66	0,52
0,959	0,60	0,49	0,093	1,58	1,05	0,393	1,13	0,80	0,612	0,64	0,51
0,394	0,92	0,79	0,078	3,48	3,29	0,409	0,96	0,69	0,631	0,62	0,48
0,241	0,61	0,49	0,012	2,28	1,78	0,379	1,03	0,73	0,626	0,63	0,49
0,976	0,75	0,61	0,098	1,32	1,13	0,424	0,94	0,69	0,547	0,69	0,54
0,501	1,32	1,00	0,123	1,04	0,84	0,419	1,07	0,77	0,596	0,65	0,51
0,618	0,76	0,63	0,177	0,84	0,67	0,393	1,01	0,71	0,624	0,63	0,49
0,507	0,81	0,66	0,098	1,32	1,13	0,439	1,04	0,76	0,624	0,63	0,49
0,745	0,98	0,84	0,086	1,19	0,94	0,431	1,08	0,79	0,616	0,64	0,49
0,728	0,61	0,48	0,078	3,48	3,29	0,399	1,00	0,74	0,588	0,66	0,51
0,394	1,23	0,93	0,001	0,90	0,73	0,410	1,01	0,72	0,582	0,66	0,51
0,525	0,80	0,65	0,098	1,32	1,13	0,397	1,03	0,73	0,618	0,64	0,50
0,278	0,73	0,59	0,170	0,95	0,78	0,449	0,96	0,71	0,605	0,64	0,50
0,488	0,61	0,49	0,064	1,73	1,52	0,401	1,04	0,75	0,586	0,66	0,52
0,394	1,73	1,63	0,098	2,47	2,33	0,381	1,01	0,71	0,591	0,65	0,51
0,394	1,73	1,63	0,128	1,64	1,06	0,399	1,05	0,74	0,622	0,64	0,49
0,671	0,62	0,51	0,098	1,32	1,13	0,435	0,96	0,70	0,610	0,64	0,49
0,505	0,88	0,74	0,193	1,14	0,91	0,431	0,95	0,68	0,608	0,64	0,50
0,192	1,09	0,79	0,206	1,20	0,95	0,427	1,01	0,75	0,600	0,65	0,51
0,486	0,90	0,73	0,125	1,40	1,03	0,462	1,18	0,89	0,595	0,65	0,51
0,218	0,60	0,48	0,098	1,32	1,13	0,409	1,04	0,76	0,614	0,64	0,50
0,394	1,73	1,63	0,098	2,47	2,33	0,431	1,02	0,75	0,621	0,63	0,50

Tabela 18 - Resultados obtidos no conjunto de dados do câncer de mama.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,150	0,84	0,68	0,729	0,48	0,45	0,863	0,50	0,17	0,944	0,30	0,08
0,937	0,58	0,51	0,176	0,65	0,52	0,844	0,58	0,28	0,943	0,31	0,08
0,456	1,35	0,99	0,908	0,57	0,49	0,809	0,60	0,27	0,961	0,26	0,05
0,568	0,80	0,66	0,687	0,81	0,63	0,881	0,53	0,31	0,949	0,28	0,07
0,916	0,62	0,56	0,795	0,56	0,46	0,871	0,51	0,23	0,946	0,30	0,09
0,150	0,84	0,68	0,830	0,60	0,46	0,852	0,53	0,18	0,949	0,29	0,06
0,755	0,80	0,66	0,542	0,61	0,50	0,862	0,55	0,27	0,960	0,26	0,05
0,654	0,62	0,51	0,357	1,33	0,87	0,865	0,55	0,26	0,952	0,28	0,09
0,461	0,86	0,70	0,456	1,35	0,99	0,864	0,59	0,38	0,950	0,28	0,07
0,708	0,73	0,59	0,861	0,48	0,43	0,881	0,45	0,17	0,938	0,33	0,06
0,456	1,35	0,99	0,470	0,50	0,44	0,876	0,54	0,24	0,952	0,28	0,07
0,872	0,62	0,56	0,391	0,95	0,80	0,888	0,47	0,21	0,949	0,28	0,06
0,342	0,65	0,58	0,357	1,33	0,87	0,860	0,51	0,17	0,935	0,34	0,09
0,007	0,82	0,67	0,106	0,74	0,56	0,869	0,52	0,25	0,953	0,27	0,05
0,004	0,82	0,67	0,376	0,53	0,47	0,865	0,53	0,20	0,950	0,27	0,07
0,906	1,07	0,84	0,897	0,52	0,45	0,856	0,54	0,22	0,958	0,27	0,06
0,268	0,67	0,55	0,456	1,35	0,99	0,829	0,60	0,21	0,949	0,30	0,07
0,191	0,83	0,67	0,357	1,33	0,87	0,902	0,44	0,16	0,964	0,24	0,06
0,326	0,91	0,73	0,720	0,71	0,51	0,821	0,59	0,21	0,951	0,28	0,06
0,781	0,68	0,60	0,731	0,65	0,48	0,858	0,51	0,21	0,953	0,29	0,09
0,902	0,62	0,55	0,656	1,06	0,79	0,774	0,72	0,26	0,942	0,30	0,06
0,470	0,74	0,62	0,196	0,62	0,47	0,863	0,50	0,22	0,940	0,31	0,06
0,778	0,95	0,78	0,215	0,81	0,61	0,864	0,63	0,43	0,947	0,30	0,07
0,385	0,79	0,66	0,323	0,49	0,44	0,861	0,53	0,28	0,950	0,30	0,07
0,250	1,04	0,85	0,758	0,72	0,55	0,884	0,47	0,17	0,951	0,27	0,07
0,102	0,84	0,68	0,880	0,73	0,58	0,799	0,62	0,24	0,931	0,35	0,09
0,420	0,75	0,61	0,908	1,03	0,78	0,820	0,59	0,18	0,952	0,29	0,08
0,381	0,97	0,56	0,656	0,62	0,53	0,884	0,46	0,17	0,948	0,29	0,09
0,456	1,35	0,99	0,290	0,56	0,49	0,895	0,44	0,20	0,952	0,27	0,05
0,215	0,83	0,68	0,432	0,48	0,45	0,842	0,55	0,21	0,942	0,31	0,10

Tabela 19 - Resultados obtidos no conjunto de dados do mofo branco.

PGB			PGC			RNAB			RNAR		
Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE	Cor.	RMSE	MAE
0,745	56,12	44,47	0,250	120,48	44,75	0,405	75,65	53,41	0,747	42,24	24,12
0,774	53,22	36,48	0,220	154,78	46,06	0,372	74,22	41,69	0,595	46,78	25,08
0,587	57,00	45,24	0,162	167,25	46,98	0,426	63,87	36,83	0,519	54,34	27,82
0,575	56,05	42,46	0,157	145,02	48,59	0,475	76,37	41,52	0,633	44,53	24,25
0,554	54,20	43,35	0,374	130,17	48,88	0,280	85,45	53,83	0,648	53,36	30,82
0,566	60,49	51,75	0,253	145,15	46,13	0,365	72,54	38,86	0,567	45,83	25,36
0,687	54,14	38,18	0,345	123,41	47,34	0,254	106,59	55,02	0,570	57,44	30,67
0,574	56,53	44,40	0,076	127,84	47,28	0,281	90,46	57,37	0,639	53,27	30,61
0,714	54,43	45,79	0,211	164,09	47,78	0,366	72,18	40,26	0,484	51,57	30,30
0,488	52,83	47,65	0,275	163,70	47,33	0,384	92,02	47,37	0,777	45,98	29,25
0,508	56,69	45,39	0,214	142,84	47,46	0,029	116,90	54,36	0,684	54,69	27,35
0,743	57,34	40,37	0,053	159,66	46,42	0,366	85,90	46,88	0,729	44,15	26,29
0,494	53,77	50,27	0,390	131,38	47,18	0,415	88,98	44,01	0,666	42,79	23,52
0,398	56,52	40,40	0,103	151,05	47,06	0,292	85,87	51,43	0,680	44,66	23,74
0,667	57,92	39,08	0,343	105,80	46,29	0,336	92,29	53,73	0,517	57,45	34,28
0,490	59,80	50,20	0,198	156,10	46,43	0,341	104,03	53,01	0,657	43,15	22,26
0,717	53,41	41,28	0,269	166,61	48,91	0,453	71,48	45,69	0,515	52,27	30,12
0,403	59,30	57,98	0,163	145,44	48,51	0,416	63,46	41,48	0,763	48,54	24,35
0,741	57,99	39,47	0,215	143,09	48,41	0,334	97,09	51,37	0,724	47,35	29,04
0,734	54,20	43,14	0,231	115,50	46,73	0,501	61,40	38,94	0,688	41,51	23,13
0,605	56,25	36,90	0,301	161,16	47,81	0,245	93,72	52,44	0,617	51,39	30,92
0,588	53,71	48,64	0,395	134,99	46,51	0,419	72,63	42,56	0,681	47,59	25,01
0,663	53,54	54,70	0,345	160,13	48,68	0,296	80,49	47,23	0,724	44,29	24,19
0,743	55,76	61,80	0,195	148,02	48,00	0,363	103,52	57,74	0,712	44,48	25,00
0,497	61,01	47,22	0,285	135,89	47,82	0,352	85,25	43,61	0,577	53,24	30,80
0,706	62,91	43,83	0,332	165,37	48,85	0,270	88,48	58,36	0,694	44,16	22,96
0,665	56,79	48,57	0,205	144,33	46,89	0,337	88,14	45,29	0,654	39,06	21,09
0,275	60,27	44,71	0,282	150,94	48,45	0,508	71,80	40,43	0,706	43,18	24,90
0,734	53,43	37,03	0,312	145,55	47,34	0,393	82,86	47,30	0,555	45,37	24,70
0,689	53,39	48,22	0,364	127,74	48,67	0,376	70,06	41,62	0,763	44,18	26,47

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
MESTRADO EM COMPUTAÇÃO APLICADA

ALISON ROGER HAJO WEBER

PROGRAMAÇÃO GENÉTICA, REDES NEURAIS ARTIFICIAIS E TÉCNICAS DE BALANCEAMENTO NA MODELAGEM DE DADOS AGRÍCOLAS: ESTUDO DA DOENÇA MOFO BRANCO

PONTA GROSSA

2012