5
Sistema Híbrido Neuro-Fuzzy-Genético para Mineração Automática de Dados

5.1.
Introdução

Após a descrição do Sistema NFHB juntamente com seus parâmetros e do modelo coevolutivo hierárquico, é necessário modelar um sistema complementar que seja capaz de otimizar esses parâmetros, descobrindo-os automaticamente sem a presença de um especialista.

Em [LANA00] foi utilizado um AG para otimizar parte dos parâmetros do sistema NFHB em aplicações de previsão de séries temporais. Nessa abordagem, porém, a presença de um especialista não foi evitada pelo fato de existirem outros parâmetros a serem definidos. Além disso, o principal objetivo da otimização realizada não era eliminar a presença de um especialista mas sim encontrar uma melhor configuração para os parâmetros mais empíricos do sistema.

Para a construção de um sistema inteiramente automático, todos os parâmetros do sistema NFHB devem ser determinados. Como foi visto no capítulo anterior, são onze os parâmetros na Tabela 1, o que torna o problema excessivamente complexo para um único AG.

Ainda neste capítulo será descrito o sistema proposto desenvolvido para otimizar os parâmetros do Sistema NFHB. Os AG's componentes e a configuração correspondente também serão detalhados, concluindo a definição completa do sistema Híbrido Neuro-Fuzzy-Genético para Mineração Automática de Dados.

5.2.
O Sistema Evolutivo de Otimização dos Parâmetros

O problema de otimização dos parâmetros do sistema NFHB é um problema complexo de otimização. Deste modo, foi proposto e desenvolvido um sistema coevolutivo hierárquico para resolvê-lo com eficácia.

PUC-Rio - Certificação Digital Nº 0210427/CA

Após uma análise detalhada dos parâmetros a serem otimizados, descritos na seção 3.8, verificou-se que esses poderiam ser agrupados em quatro diferentes grupos: parâmetros de configuração do sistema NFHB; parâmetros relacionados aos antecedentes das regras; parâmetros relacionados aos conseqüentes das regras; e parâmetros específicos do problema em questão.

Neste contexto, foi criada uma hierarquia entre esses grupos, onde os parâmetros de configuração do sistema NFHB ficam no topo da hierarquia e referenciam parâmetros relacionados a antecedentes, conseqüentes e específicos do problema, em um nível abaixo.

A Figura 23 representa a hierarquia coevolutiva criada.

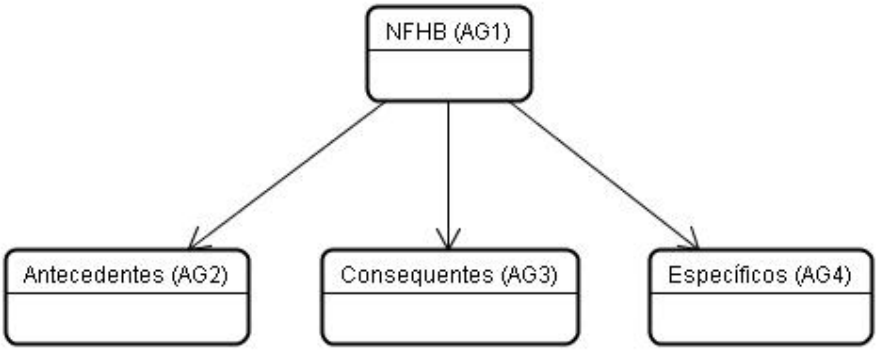

Figura 23: Hierarquia de Parâmetros do Sistema NFHB

Dessa forma, é possível definir quais parâmetros do sistema se encaixam em cada um dos grupos criados. O agrupamento é mostrado na Figura 24.

NFHB
δ
MétodoSeleçãoVariáveis
UsarTolerânciaSeparação

Antecedentes
$\dot{a}$
$\dot{x}$
$\tilde{x}$
numCiclosRProp
step

Consequentes
NívelCombinaçãoLinear
numCiclosMQO
UsarCombinaçãoLinear

Figura 24: Parâmetros do Sistema NFHB Agrupados

O grupo de parâmetros específicos depende do problema a ser resolvido (previsão, classificação etc), deixando o sistema flexível para ser usado em diferentes aplicações. No caso de previsão de séries temporais, existe um parâmetro específico do problema a ser otimizado, a janela de previsão. Para o caso de classificação de padrões, não existem parâmetros específicos do problema e, conseqüentemente, não existirá um AG correspondente na

PUC-Rio - Certificação Digital Nº 0210427/CA

coevolução. Outros problemas podem possuir parâmetros passíveis de otimização e utilizar o AG correspondente.

1) As tabelas de parâmetros apresentadas na Figura 24 também representam como são formados os cromossomas que evoluem nas espécies criadas. O cromossoma do AG1 não guarda referência a indivíduos das populações inferiores como genes. No momento do cálculo das avaliações da espécie de nível superior, um indivíduo de cada população do nível inferior é selecionado para cada indivíduo da população principal, formando conjuntos completos de parâmetros do Sistema NFHB a serem avaliados. O diagrama de atividades que representa o algoritmo de seleção de indivíduos das populações de nível inferior do sistema modelado está representado na Figura 25.

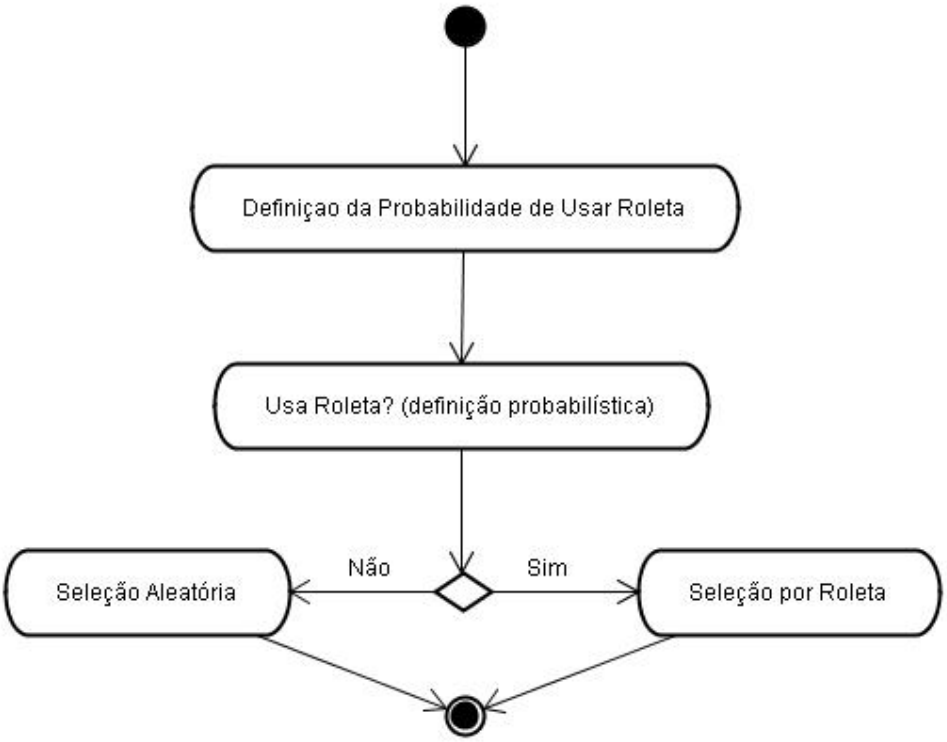

Figura 25: Diagrama de Atividade da Seleção de Indivíduos das Populações Inferiores

A seleção de indivíduos das populações de nível inferior é realizada seguindo as seguintes regras:

2) É especificada *a priori* uma probabilidade, chamada pRoleta, que estabelece a freqüência com que indivíduos serão escolhidos através de um mecanismo de roleta;

PUC-Rio - Certificação Digital Nº 0210427/CA

3) Escolhe-se probabilisticamente se será utilizado um mecanismo de roleta;
4) Se for escolhido o mecanismo de roleta, esse método é utilizado para selecionar os indivíduos que contribuirão com seus cromossomas. Os mais aptos têm maior chance de serem escolhidos, pois a roleta leva em consideração esse critério;
5) Se não foi selecionado o mecanismo de roleta, os indivíduos são selecionados aleatoriamente para que todos tenham chances de participar.

O sincronismo entre os AGs segue o mesmo modelo definido na seção anterior. As funções de avaliação e os operadores genéticos utilizados pelos AGs serão descritos nas próximas seções.

PUC-Rio - Certificação Digital Nº 0210427/CA

5.2.1. Funções de Avaliação

A definição das funções de avaliação dos AGs componentes é um fator essencial para o resultado final do modelo. Uma função de avaliação mal formulada pode fazer com que a espécie não evolua por não favorecer os melhores indivíduos de acordo com o objetivo desejado.

O cálculo das avaliações dos AGs do segundo nível é feito da mesma maneira para todos os AGs, apesar de isso não ser uma regra. Como foi dito, esse cálculo depende dos valores das avaliações provenientes dos indivíduos da população pai e é realizado da seguinte forma.

$$f_{i(j)} = média\ (f_{1(b)},..., f_{1(d)}) \quad \text{Eq. 42}$$

Onde:

- $f_{i(j)}$ é o valor da avaliação do indivíduo *j*, pertencente a população do AG *i* do nível inferior na hierarquia;
- $f_{1(b)}$ é o valor da avaliação do indivíduo *b*, pertencente ao AG principal;
- $f_{1(d)}$ é o valor da avaliação do indivíduo *d*, pertencente ao AG principal.

Como pode ser visto na equação 42, a avaliação de um indivíduo de uma população filha é a média aritmética dos valores das avaliações dos indivíduos

da população pai que o utilizaram em suas avaliações. No exemplo demonstrado na equação, os indivíduos *b* e *d* da população principal selecionaram e utilizaram o indivíduo *j* da população filha em questão.

A avaliação leva em conta a participação dos indivíduos em todas as vezes que são selecionados e, sendo assim, os indivíduos que, na média, contribuírem mais para um menor erro do sistema, serão privilegiados na evolução.

Como foi visto no capítulo anterior, para o cálculo da avaliação de um indivíduo do AG principal é necessário selecionar um representante de cada AG filho, formando um conjunto completo de valores para os parâmetros do sistema NFHB. A partir desses valores, o sistema pode ser configurado e treinado. Ao longo do treinamento, métricas de avaliação do sistema referentes ao problema em questão são calculadas para posteriormente serem utilizadas no cálculo efetivo do valor da avaliação dos indivíduos do AG principal. A seguir serão descritas as funções de avaliação dos AGs principais para as aplicações de previsão de séries temporais e classificação de padrões.

5.2.1.1. Previsão de Séries Temporais

Para aplicações de previsão de séries temporais as métricas mais utilizadas para avaliação de eficácia de modelos são: MAPE (*Mean Absolute Percentage Error*), RMSE (*Root Mean Square Error*) e Utheil.

O MAPE é a média percentual dos erros ao longo da validação do sistema e é representado pela seguinte equação.

$$MAPE = \frac{\sum_{k=1}^{N} \left| \frac{a_k - y_k}{a_k} \right|}{N} \cdot 100\% \qquad \text{Eq. 43}$$

Onde:

- a_k é a saída esperada para o padrão *k*;
- y_k é a saída prevista ou calculada pelo sistema para o padrão *k*;
- N é o número total de previsões.

O RMSE é o erro médio quadrático das previsões representado pela seguinte fórmula.

PUC-Rio - Certificação Digital Nº 0210427/CA

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} |a_k - y_k|^2}{N}} \qquad \text{Eq. 44}$$

Este erro penaliza os erros maiores pelo fato de os elevar ao quadrado, diferentemente do MAPE.

Outra métrica aplicada à previsão de séries temporais é o Utheil. Essa métrica testa se o algoritmo de previsão obtém um resultado melhor do que a previsão ingênua, ou seja, sempre prever o instante seguinte como o valor anterior na série, onde

$$U = \frac{\sqrt{\sum_{k=1}^{N} (a_k - y_k)^2}}{\sqrt{\sum (a_k - a_{k-1})^2}} \qquad \text{Eq. 45}$$

A partir das definições dessas métricas, a função de avaliação proposta para o AG principal em aplicações de previsão de séries temporais é a seguinte.

```
Se  (Utheil_Validação >= 1.0 ) Então
{
  Avaliação = Utheil_Validação;
}
Senão
{
  Avaliação = (MapeValidação)^2 * Utheil_Validação;
}
End.
```

Como pode ser visto, a função de avaliação penaliza os indivíduos que possuem Utheil maior do que um, já que são piores que a previsão ingênua. Caso o Utheil seja menor do que um, é feita uma avaliação que prioriza o MAPE, elevando-o ao quadrado. O erro RMSE não é utilizado na avaliação pelo fato de não ser padronizado e estar em uma escala diferente das outras métricas.

O AG objetiva evoluir indivíduos que minimizem o valor gerado pela função de avaliação.

PUC-Rio - Certificação Digital Nº 0210427/CA

5.2.1.2. Classificação de Padrões

No caso de classificação de padrões, a métrica mais utilizada é a porcentagem de acerto do modelo, ou seja, a proporção de acertos de classificação dentre o número total de padrões avaliados. A equação que define esse cálculo é

$$P = \frac{A}{N} \quad \text{Eq. 46}$$

Onde:

- A é o número de acertos de classificação, ou seja, quantas vezes o sistema acerta exatamente a classe a que pertence um padrão;
- N é o número total de padrões classificados.

Desse modo, a função de avaliação proposta para o AG principal em aplicações de classificação de padrões é a seguinte.

$$Avaliação = (MAPE\_Validação)^2 / P\_Validação \quad \text{Eq. 47}$$

Nota-se que nessa função estão sendo dadas importâncias iguais ao MAPE de validação e ao percentual de acerto de validação, porém deseja-se minimizar o MAPE e maximizar o percentual. Novamente os indivíduos de menor valor de avaliação serão beneficiados na evolução.

5.2.2. Configuração dos Algoritmos Genéticos

Após a definição da arquitetura coevolutiva hierárquica e das funções de avaliação dos AGs componentes, é necessário definir a configuração dos mesmos.

Todos os AGs utilizam o mecanismo de *SteadyState*, responsável por preservar os melhores indivíduos da população, garantindo que o material genético desses indivíduos será mantido para a próxima geração da evolução. O número de indivíduos que permanecem evoluindo é definido de forma percentual por um parâmetro chamado *pSteadyState*.

A seleção de indivíduos para contribuir na geração da nova população é realizada por um mecanismo de roleta, onde são consideradas as avaliações

PUC-Rio - Certificação Digital Nº 0210427/CA

dos indivíduos, ou seja, indivíduos mais aptos possuem maior probabilidade de serem utilizados.

Após a seleção, são realizadas operações de *crossover* e mutação na geração de novos indivíduos para compor a nova população.

Conforme descrito no capítulo 2, a operação de *crossover* objetiva realizar a troca de material genético entre indivíduos da população. Esse operador possui um parâmetro que define a probabilidade do mesmo ser utilizado. O teste que define se o operador deve ser utilizado ou não ocorre *N* vezes, onde *N* é o número de indivíduos da população.

Para cada teste, são escolhidos aleatoriamente dois indivíduos como candidatos a troca de material genético. Depois de escolhidos os candidatos, o teste probabilístico é realizado, definindo se o operador é aplicado ou não. Caso o teste seja positivo, o *crossover* é realizado.

Os cromossomas de todos os AGs possuem apenas genes do tipo real e, conseqüentemente, devem ser utilizados operadores correspondentes a esse tipo de gene. Vários tipos de *crossover* são utilizados: aritmético, simples, geométrico, esférico e heurístico [MICH94]. Todos esses tipos de *crossover* são configurados para ocorrer com a mesma freqüência e são selecionados aleatoriamente.

A operação de mutação é responsável por gerar variações no material genético da população explorando regiões do domínio do problema ainda inexploradas. Da mesma forma que no *crossover*, esta operação possui um parâmetro que define a probabilidade de ocorrência.

O mecanismo de aplicação do operador é semelhante ao utilizado no *crossover*, porém a mutação faz sentido ser aplicada apenas a um indivíduo. Sendo assim, são escolhidos *N* indivíduos aleatoriamente que serão modificados ou não probabilisticamente.

Também são utilizados vários tipos de mutação aplicados a genomas com valores reais, são eles: uniforme, limite, não-uniforme e gaussiano [MICH94]. Todos esses tipos de mutação também são configurados para ocorrer com a mesma freqüência e são selecionados aleatoriamente.

Todos os AGs podem rodar por vários ciclos com várias gerações, sendo esses valores correspondentes a dois parâmetros do sistema. Entre dois ciclos subseqüentes, um percentual de indivíduos pode permanecer inalterado, preservando uma carga genética com boa avaliação. Esse percentual é definido pelo parâmetro *pPersistence*.

PUC-Rio - Certificação Digital Nº 0210427/CA

Além disso, são configurados o número de indivíduos das populações, a probabilidade de ser utilizado o mecanismo de roleta pelo AG principal na seleção de representantes (ver seção 4.2.2) e ainda se o AG deve beneficiar indivíduos com baixa avaliação na evolução. Na Tabela 2 é mostrado um resumo dos parâmetros de configuração dos AGs.

Tabela 2: Tabela de Parâmetros de Configuração dos AGs

Nome	Tipo	Função
Ciclos	Inteiro	Define o número de ciclos que o AG irá evoluir.
Gerações	Inteiro	Número de gerações em um ciclo de evolução.
Genomas	Inteiro	Número de indivíduos da população.
pMutação	Real	Probabilidade de ser realizada uma mutação genética em um indivíduo da população.
pCrossover	Real	Probabilidade de ser realizado crossover entre dois indivíduos da população.
pSteadyState	Real	Proporção de melhores indivíduos que permanecem na população após uma geração.
pRoleta (AG Principal)	Real	Probabilidade de ser utilizado o mecanismo de roleta na seleção de indivíduos de populações de níveis inferiores (ver seção 4.2.2).
Minimização	Booleano	Indica se o AG deve priorizar indivíduos com baixa avaliação.
pPersistence	Real	Porcentual de melhores indivíduos que irão permanecer de ciclo em ciclo.
nGenes	Real	Número de genes no cromossoma dos indivíduos do AG em questão.

Estes parâmetros devem ser especificados *a priori* em todos os AGs para que o sistema como um todo continue sendo automático. Dessa forma, esses parâmetros foram definidos empiricamente.

Os valores definidos para os parâmetros dos AGs correspondentes ao sistema NFHB, aos antecedentes e aos conseqüentes são descritos na Tabela 3.

PUC-Rio - Certificação Digital Nº 0210427/CA

Tabela 3: Tabela de Valores dos Parâmetros de Configuração dos AG's

Nome	AG1	AG2	AG3
Ciclos	3	3	3
Gerações	80	80	80
Genomas	50	50	40
pMutação	0.2	0.15	0.15
pCrossover	0.6	0.6	0.6
pSteadyState	0.3	0.4	0.4
pRoleta	0.4	-	-
Minimização	true	true	true
pPersistence	0.1	0.1	0.1
nGenes	3	5	3

Para se chegar a esses valores, várias outras configurações foram testadas. Tentou-se, inicialmente, configurar todos os AGs com os mesmos valores para todos os parâmetros. Os resultados melhoraram ao se aumentar a probabilidade de mutação do AG principal, fazendo com que houvesse uma maior variação genética, evitando que a evolução ficasse em torno de mínimos locais.

Foram obtidos melhores resultados modificando o parâmetro *pSteadyState* na população principal, baixando o seu valor para 0.3, diminuindo um pouco a proporção de indivíduos que permanecem de geração para geração, dando uma chance maior de surgirem novos indivíduos com boa aptidão. Foram utilizados mais indivíduos nos AGs 1 e 2 porque esse evoluem um maior número de parâmetros e, conseqüentemente, necessitam de maior diversidade genética.

Tentou-se aumentar o valor do parâmetro *pRoleta* na população principal, porém os resultados pioraram por não serem dadas chances a indivíduos ainda sem nenhuma avaliação e que muitas vezes poderiam ser bons. Percebe-se que as populações filhas não possuem valores para esse parâmetro já que essas populações não possuem filhos a serem selecionados.

Cada parâmetro sendo otimizado corresponde a um gene em um dos cromossomas. Para cada gene deve ser definido um domínio de valores mínimo e máximo, definindo o espaço de busca dos valores. Assim, foram especificados domínios para todos os parâmetros em questão, os quais estão detalhados na Tabela 4.

PUC-Rio - Certificação Digital Nº 0210427/CA

Tabela 4: Domínios dos Parâmetros do Sistema NFHB

Parâmetro	Mínimo	Máximo
δ	0.09	0.2
MétodoSeleçãoVariáveis	0.001	3
UsarTolerânciaSeparação	0	1
a^*	1	6
λ^+	1.1	1.5
λ^-	0.1	0.5
numCiclosRProp	10	40
step	50	120
NívelCombinaçãoLinear	0	10
numCiclosMQO	20	80
UsarCombinaçãoLinear	0	1

Os valores mínimo e máximo dos parâmetros foram definidos após um estudo dos valores definidos por especialistas. Os valores foram escolhidos de forma que os domínios englobassem os melhores valores para cada um dos parâmetros do sistema NFHB em todas as aplicações.

5.2.2.1. Previsão de Séries Temporais

Como já foi dito anteriormente, existe um AG componente da coevolução responsável por evoluir parâmetros específicos do problema sendo resolvido. No caso de previsão de séries temporais, é criado um AG para evoluir a janela de previsão.

A melhor configuração encontrada para o AG de otimização da janela de previsão está definida na Tabela 5.

Tabela 5: Tabela de Valores dos Parâmetros de Configuração do AG4

Nome	AG4
Ciclos	3
Gerações	80
Genomas	30
pMutação	0.15
pCrossover	0.6
pSteadyState	0.4
pRoleta	-
Minimização	true
pPersistence	0.1
nGenes	1

PUC-Rio - Certificação Digital Nº 0210427/CA

Foram utilizados poucos genomas nessa população já que está se otimizando apenas um parâmetro, não sendo necessária uma população maior.

O domínio da janela de previsão deve ser especificado pelo usuário pois depende do problema de previsão em questão. No caso em que os valores de previsão são mensais, um bom domínio são janelas entre 4 e 12. Neste trabalho foram utilizados esses valores, já que foram testadas séries mensais de carga elétrica. Para outros problemas de previsão devem ser definidos valores coerentes conforme a amostragem dos dados.

Todas as descrições realizadas na seção anterior para os outros AGs são válidas para o AG de parâmetros específicos.

No próximo capítulo serão descritos os estudos de caso, com os resultados e as análises correspondentes.

PUC-Rio - Certificação Digital Nº 0210427/CA