



**Manoel Roberto Aguirre de Almeida**

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

**Dissertação de Mestrado**

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientadores:

Marley M. B. R. Velasco  
Marco Aurélio C. Pacheco

Rio de Janeiro, fevereiro de 2004



**Manoel Roberto Aguirre de Almeida**

## **Sistema Híbrido Neuro-Fuzzy-Genético para Mineração Automática de Dados**

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

**Prof. Marco Aurélio C. Pacheco**  
Orientador

**Prof. Ricardo Tanscheit**  
PUC-Rio

**Prof<sup>a</sup>. Karla T. Figueiredo Leite**  
PUC-Rio

**Prof. Valmir C. Barbosa**  
UFRJ

**Prof. Leandro dos Santos Coelho**  
PUC-PR

**Prof. Flávio Joaquim de Souza**  
UERJ

**Prof. José Eugênio Leal**  
Coordenador Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 25 de março de 2004

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

**Manoel Roberto Aguirre de Almeida**

Graduou-se em Engenharia de Computação pela PUC-Rio em Julho de 2001. Iniciou seus estudos de mestrado na área de Métodos de Apoio à Decisão no Departamento de Elétrica em Fevereiro de 2002.

Ficha Catalográfica

Almeida, Manoel Roberto Aguirre de

Sistema híbrido neuro-fuzzy-genético para mineração automática de dados / Manoel Roberto Aguirre de Almeida ; orientadores: Marley M. B. R. Velasco, Marco Aurélio C. Pacheco. – Rio de Janeiro : PUC, Departamento de Engenharia Elétrica, 2004.

112 f. : il. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Incluí referências bibliográficas.

1. Engenharia elétrica – Teses. 2. Mineração de dados. 3. Sistemas neuro-fuzzy. 4. Coevolução genética. I. Velasco, Marley M. B. R. II. Pacheco, Marco Aurélio C. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

À minha família

## Agradecimentos

A Deus por ter me permitido existir e ter me guiado nesta jornada.

Ao CNPq, pelo apoio financeiro.

À Profa. Dra. Marley Maria B. R. Vellasco e ao Prof. Dr. Marco Aurélio C. Pacheco, orientadores desta tese, pelo apoio, carinho, incentivo e confiança depositados.

Ao meu pai por não medir esforços para me dar a melhor educação e me apoiar em todos os momentos do desenvolvimento deste trabalho.

À minha mãe que sempre sonhou em presenciar este momento e que, com certeza, em algum lugar está muito feliz por minha conquista.

Ao meu avô José e meus tios por acreditarem em mim e estarem sempre presentes com seu carinho.

À Raquel Andrade, pelo carinho, confiança, amor e, principalmente, pela paciência ao longo da elaboração desta dissertação.

Aos amigos João Machado e Danilo Tuler pela amizade e companheirismo nos momentos de sufoco.

Aos amigos do ICA com quem sempre pude contar e tirar dúvidas fundamentais no processo de pesquisa.

À Pontifícia Universidade Católica do Rio de Janeiro.

## Resumo

Manoel Roberto Aguirre de Almeida. **Sistema Híbrido Neuro-Fuzzy-Genético para Mineração Automática de Dados**. Rio de Janeiro, 2004. 112p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação apresenta a proposta e o desenvolvimento de um sistema de mineração de dados inteiramente automático. O objetivo principal é criar um sistema que seja capaz de realizar a extração de informações obscuras a partir de bases de dados complexas, sem exigir a presença de um especialista técnico para configurá-lo. O sistema híbrido neuro-fuzzy hierárquico com particionamento binário (NFHB) vem apresentando excelentes resultados em tarefas de classificação de padrões e previsão, além de possuir importantes características não encontradas em outros sistemas similares, entre elas: aprendizado automático de sua estrutura; capacidade de receber um número maior de entradas abrangendo um maior número de aplicações; e geração de regras lingüísticas como produto de seu treinamento. Entretanto, este modelo ainda necessita de uma complexa parametrização inicial antes de seu treinamento, impedindo que o processo seja automático em sua totalidade. O novo modelo proposto busca otimizar a parametrização do sistema NFHB utilizando a técnica de coevolução genética, criando assim um novo sistema de mineração de dados completamente automático. O trabalho foi realizado em quatro partes principais: avaliação de sistemas existentes utilizados na mineração de dados; estudo do sistema NFHB e a determinação de seus principais parâmetros; desenvolvimento do sistema híbrido neuro-fuzzy-genético automático para mineração de dados; e o estudo de casos.

No estudo dos sistemas existentes para mineração de dados buscou-se encontrar algum modelo que apresentasse bons resultados e ainda fosse passível de automatização. Várias técnicas foram estudadas, entre elas: Métodos Estatísticos, Árvores de Decisão, Associação de Regras, Algoritmos Genéticos, Redes Neurais Artificiais, Sistemas Fuzzy e Sistemas Neuro-Fuzzy. O sistema NFHB foi escolhido como sistema de inferência e extração de regras para a realização da mineração de dados. Deste modo, este modelo foi estudado e seus parâmetros mais importantes foram determinados. Além disso, técnicas de seleção de variáveis de entradas foram investigadas para servirem como opções para o modelo. Ao final, foi obtido um conjunto de parâmetros que deve ser automaticamente determinado para a completa configuração deste sistema.

Um modelo coevolutivo genético hierárquico foi criado para realizar com excelência a tarefa de otimização do sistema NFHB. Desta forma, foi modelada uma arquitetura hierárquica de Algoritmos Genéticos (AG's), onde os mesmos realizam tarefas de otimização complementares. Nesta etapa, também foram determinados os melhores operadores genéticos, a parametrização dos AG's, a melhor representação dos cromossomas e as funções de avaliação. O melhor conjunto de parâmetros encontrado é utilizado na configuração do NFHB, tornando o processo inteiramente automático.

No estudo de casos, vários testes foram realizados em bases de dados reais e do tipo benchmark. Para problemas de previsão, foram utilizadas séries de carga de energia elétrica de seis empresas: Cerj, Copel, Eletropaulo, Cemig, Furnas e Light. Na área de classificação de padrões, foram utilizadas bases conhecidas de vários artigos da área como *Glass Data*, *Wine Data*, *Bupa Liver Disorders* e *Pima Indian Diabetes*. Após a realização dos testes, foi feita uma comparação com os resultados obtidos por vários algoritmos e pelo NFHB original, porém com parâmetros determinados por um especialista.

Os testes mostraram que o modelo criado obteve resultados bastante satisfatórios, pois foi possível, com um processo completamente automático, obter taxas de erro semelhantes às obtidas por um especialista, e em alguns casos taxas menores. Desta forma, um usuário do sistema, sem qualquer conhecimento técnico sobre os modelos utilizados, pode utilizá-lo para realizar mineração de bancos de dados, extraindo informações e até mesmo conhecimento que podem auxiliá-lo em processos de tomada de decisão, o qual é o objetivo final de um processo de Knowledge Data Discovery.

## **Palavras-chave**

Mineração de Dados; Sistemas Neuro-Fuzzy; Coevolução Genética

## Abstract

Manoel Roberto Aguirre de Almeida. **Híbrido Neuro-Fuzzy-Genético para Mineração de Dados Automática**. Rio de Janeiro, 2004. 112p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation presents the proposal and the development of a totally automatic data mining system. The main objective is to create a system that is capable of extracting obscure information from complex databases, without demanding the presence of a technical specialist to configure it. The Hierarchical Neuro-Fuzzy Binary Space Partitioning model (NFHB) has produced excellent results in pattern classification and time series forecasting tasks. Additionally, it provides important features that are not present in other similar systems, such as: automatic learning of its structure; ability to deal with a larger number of input variables, thus increasing the range of possible applications; and generation of linguistic rules as a result of its training process. However, this model depends on a complex configuration process before the training is performed, hindering to achieve a totally automatic system. The model proposed in this Dissertation tries to optimize the NFHB system parameters by using the genetic coevolution technique, thus creating a new automatic data mining system. This work consisted of four main parts: evaluation of existing systems used in data mining; study of the NFHB system and definition of its main parameters; development of the automatic hybrid neuro-fuzzy-genetic system for data mining; and case studies.

In the study of existing data mining systems, the aim was to find a suitable model that could yield good results and still be automated. Several techniques have been studied, among them: Statistical methods, Decision Trees, Rules Association, Genetic Algorithms, Artificial Neural Networks, Fuzzy and Neuro-Fuzzy Systems. The NFHB System was chosen for inference and rule extraction in the data mining process. In this way, this model was carefully studied and its most important parameters were determined. Moreover, input variable selection techniques were investigated, to be used with the proposed model. Finally, a set of parameters was defined, which must be determined automatically for the complete system configuration.

A hierarchical coevolutionary genetic model was created to execute the system optimization task with efficiency. Therefore, a hierarchical architecture of



genetic algorithms (GAs) was created, where the GAs execute complementary optimization tasks. In this stage, the best genetic operators, the GAs configuration, the chromosomes representation, and evaluation functions were also determined. The best set of parameters found was used in the NFHB configuration, making the process entirely automatic.

In the case studies, various tests were performed with benchmark databases. For forecasting problems, six electric load series were used: Cerj, Copel, Eletropaulo, Cemig, Furnas and Light. In the pattern classification area, some well known databases were used, namely Glass Data, Wine Data, Bupa Liver Disorders and Pima Indian Diabetes. After the tests were carried out, a comparison was made with known models and with the original NFHB System, configured by a specialist.

The tests have demonstrated that the proposed model generates satisfactory results, producing, with an automatic process, similar errors to the ones obtained with a specialist configuration, and, in some cases, even better results can be obtained. Therefore, a user without any technical knowledge of the system, can use it to perform data mining, extracting information and knowledge that can help him/her in decision taking processes, which is the final objective of a Knowledge Data Discovery process.

## **Keywords**

Data Mining; Neuro-Fuzzy Systems; Genetic Coevolution

## Sumário

1 Introdução	17
1.1. Motivação	17
1.2. Objetivos	18
1.3. Descrição do Trabalho	19
1.4. Organização da Tese	21
2 Sistemas de Mineração de Dados	22
2.1. Introdução	22
2.2. Classificação de Padrões	24
2.3. Previsão de Séries Temporais	24
2.4. Técnicas de Mineração de Dados	25
2.4.1. Métodos Estatísticos	25
2.4.1.1. Discriminante Linear de Fisher	25
2.4.1.2. Classificadores Bayesianos	26
2.4.1.3. Redes Bayesianas	27
2.4.2. Árvores de Decisão	27
2.4.2.1. ID3 e C4.5	28
2.4.2.2. CART ( <i>Classification And Regression Trees</i> )	29
2.4.2.3. FID3.1	29
2.4.3. Regras de Associação	30
2.4.4. Algoritmos Genéticos	30
2.4.4.1. GA-Miner	31
2.4.4.2. Rule Evolver	31
2.4.5. Redes Neurais Artificiais	31
2.4.6. Sistemas Neuro-Fuzzy	33
2.4.6.1. Sistema ANFIS	34
2.4.6.2. FSOM	36
2.4.6.3. NEFCLASS	37
3 Sistema Neuro-Fuzzy Hierárquico BSP (NFHB)	39
3.1. Introdução	39

3.2. O Particionamento BSP	40
3.3. A Célula Básica NFHB	41
3.4. A Arquitetura NFHB	45
3.5. O Algoritmo de Aprendizado	47
3.6. Técnicas para Atualização dos Pesos Fuzzy	52
3.6.1. Método de Gauss-Seidel	53
3.6.2. Método Resilient Back Propagation (RProp)	55
3.7. Estratégias para Seleção de Variáveis de Entrada	58
3.7.1. Método SIE	59
3.7.2. Método LSE	61
3.7.3. Método do Modelo Adaptado Anfis	61
3.8. Resumo dos Parâmetros do Sistema NFHB	62
4 Sistemas Coevolutivos	64
4.1. Introdução	64
4.2. O Modelo Coevolutivo Tradicional	65
4.3. Arquitetura Coevolutiva Hierárquica	68
5 Sistema Híbrido Neuro-Fuzzy-Genético para Mineração Automática de Dados	72
5.1. Introdução	72
5.2. O Sistema Evolutivo de Otimização dos Parâmetros	72
5.2.1. Funções de Avaliação	75
5.2.1.1. Previsão de Séries Temporais	76
5.2.1.2. Classificação de Padrões	78
5.2.2. Configuração dos Algoritmos Genéticos	78
5.2.2.1. Previsão de Séries Temporais	82
6 Estudo de Casos	84
6.1. Descrição	84
6.2. Previsão de Séries Temporais	85
6.2.1. Séries de Carga Elétrica	85
6.3. Classificação de Padrões	94
6.3.1. Base de Dados <i>Glass Data</i>	94
6.3.2. Base de Dados <i>Wine Data</i>	96
6.3.3. Base de Dados <i>Pima Indians Diabetes</i>	97
6.3.4. Base de Dados <i>Bupa Liver Disorders</i>	98
6.3.5. Comentários sobre os Resultados dos Parâmetros	100

6.4. Comparação com Resultados sem Otimização	101
7 Conclusões e Trabalhos Futuros	105
7.1. Conclusões	105
7.2. Trabalhos Futuros	106
8 Referências Bibliográficas	108

## Lista de figuras

Figura 1: Fases do Processo de KDD	22
Figura 2: Exemplo de uma Árvore de Decisão	28
Figura 3: Exemplo de Topologia de Uma Rede Neural Artificial Perceptron Multicamadas	32
Figura 4: Exemplo de Arquitetura de um Sistema ANFIS	35
Figura 5: Exemplo de Arquitetura de um Sistema FSOM	36
Figura 6: Exemplo de Arquitetura do Sistema NEFCLASS	38
Figura 7: Tipos Comuns de Particionamento dos Sistemas Neuro-Fuzzy	39
Figura 8: (a) Exemplo de Particionamento BSP; (b) Árvore do Particionamento	41
Figura 9: Esquema Simplificado de uma Célula Básica NFHB	42
Figura 10: Interior de uma Célula Básica NFHB	42
Figura 11: Formato das Funções de Pertinência da Célula NFHB	43
Figura 12: Célula NFHB na forma de uma Rede Neuro-Fuzzy	44
Figura 13: Exemplo de arquitetura NFHB	45
Figura 14: Particionamento Correspondente a Estrutura do Exemplo	46
Figura 15: Árvore de Particionamentos	46
Figura 16: Algoritmo de Aprendizado do Sistema NFHB	48
Figura 17: Gráfico de Exemplo de Erros de Treinamento e Validação	50
Figura 18: Mecanismo de particionamento com uso de combinação linear das entradas	52
Figura 19: Exemplo de Arquitetura Coevolutiva Tradicional com 3 AG's	66
Figura 20: Diagrama de Seqüência de uma Arquitetura Coevolutiva Tradicional	67
Figura 21: Exemplo de Arquitetura Coevolutiva Hierárquica com 4 AG's e 3 Níveis	69
Figura 22: Diagrama de Seqüência de uma Arquitetura Coevolutiva Hierárquica	71
Figura 23: Hierarquia de Parâmetros do Sistema NFHB	73
Figura 24: Parâmetros do Sistema NFHB Agrupados	73
Figura 25: Diagrama de Atividade da Seleção de Indivíduos das Populações Inferiores	74

Figura 26: Séries real e prevista da Copel	90
Figura 27: Séries real e prevista da Cemig	91
Figura 28: Séries real e prevista da Light	91
Figura 29: Séries real e prevista de Furnas	92
Figura 30: Séries real e prevista da Cerj	93
Figura 31: Séries real e prevista da Eletropaulo	93
Figura 32: Exemplo de Codificação da Saída do NFHB para Classificação	94

## Lista de tabelas

Tabela 1: Resumo dos Parâmetros de Entrada do Sistema NFHB	63
Tabela 2: Tabela de Parâmetros de Configuração dos AGs	80
Tabela 3: Tabela de Valores dos Parâmetros de Configuração dos AG's	81
Tabela 4: Domínios dos Parâmetros do Sistema NFHB	82
Tabela 5: Tabela de Valores dos Parâmetros de Configuração do AG4	82
Tabela 6: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série de Furnas	85
Tabela 7: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série da Light	86
Tabela 8: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série de Copel	86
Tabela 9: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série de Eletropaulo	87
Tabela 10: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série da Cerj	87
Tabela 11: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a série da Cemig	87
Tabela 12: Tabela de parâmetros otimizados pelo Sistema Coevolutivo para todas as séries de carga.	88
Tabela 13: Tabela de métricas de avaliação para todas as séries de carga analisadas	89
Tabela 14: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a base <i>Glass Data</i>	95
Tabela 15: Parâmetros Otimizados para a base <i>Glass Data</i>	95
Tabela 16: Métricas do Sistema NFHB Otimizado para a base <i>Glass Data</i>	96
Tabela 17: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a base <i>Wine Data</i>	96
Tabela 18: Parâmetros Otimizados para a base <i>Wine Data</i>	97
Tabela 19: Métricas do Sistema NFHB Otimizado para a base <i>Wine Data</i>	97
Tabela 20: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a base Pima Indians Diabetes	98
Tabela 21: Parâmetros Otimizados para a base Pima Indians Diabetes	98

Tabela 22: Métricas do Sistema NFHB Otimizado para a base Pima Indians Diabetes	98
Tabela 23: Seqüências de entradas escolhidas pelos métodos Anfis, LSE e SIE para a base Bupa Liver Disorders	99
Tabela 24: Parâmetros Otimizados para a base Bupa Liver Disorders	99
Tabela 25: Métricas do Sistema NFHB Otimizado para a base Bupa Liver Disorders	100
Tabela 26: Tabela de parâmetros otimizados pelo Sistema Coevolutivo para todas as bases de classificação estudadas	100
Tabela 27: Tabela Comparativa entre o modelo NFHB configurado por especialista e o NFHB Otimizado por Coevolução em aplicações de previsão	101
Tabela 28: Tabela comparativa do Modelo Coevolutivo com outros métodos de previsão	102
Tabela 29: Tabela comparativa do Modelo Coevolutivo com outros métodos de previsão	102
Tabela 30: Tabela Comparativa entre o modelo NFHB configurado por especialista e o NFHB Otimizado por Coevolução em aplicações de classificação	103
Tabela 31: Tabela comparativa do Modelo Coevolutivo com outros métodos de Classificação	104