PUC
RIO

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

Pontifícia Universidade Católica
do Rio de Janeiro

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

Profª. 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

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. **Hibrid Neuro-Fuzzy-Genetic System for Automatic Data Mining.** 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 coevolutive 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 chromossomes 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

Lista de figuras

Lista de tabelas