1
Introdução

1.1.
Motivação

Uma das áreas da Tecnologia da Informação (TI) que mais cresce atualmente é a chamada *Business Intelligence* (BI) ou Inteligência de Negócio. O termo foi criado em 1989 por Howard Dresner [POWE03], analista do Gartner Group, com o objetivo de definir uma nova gama de soluções de *software* que buscam, a partir de bancos de dados, extrair conhecimento para dar suporte a decisões estratégicas das empresas.

O conceito de sistemas de suporte à decisão não é novo. Em 1964, Michael S. Scott Morton [POWE03] já descrevia os denominados *decision support systems*, trabalho que serviu de base para os sistemas atuais de BI.

É importante identificar o que vem a ser um verdadeiro sistema de BI. Existem vários conceitos aparentemente superpostos envolvidos nessa questão, entre eles: *dados*, *informação*, *conhecimento* e *inteligência*.

Dados podem ser considerados registros em bancos de dados, os quais podem estar espalhados pelos departamentos de uma empresa. Informação é a junção de diversos dados que, em conjunto, fazem algum sentido para entendimento. Conhecimento é a informação sendo interpretada, muitas vezes em conjunto com outras informações e com experiência já adquirida, tornando-se fruto de um processo de aprendizado. Finalmente, neste contexto, inteligência pode ser definida como a capacidade de utilizar corretamente o conhecimento adquirido de forma a tomar decisões estratégicas ou de negócio.

As ferramentas que realmente realizam o *Business Intelligence* são aquelas que conseguem chegar no nível de inteligência, ou seja, extraem informações, contextualizam e adquirem conhecimento, para em seguida indicar o caminho mais "inteligente" a ser trilhado em uma decisão estratégica. As grandes empresas buscam ferramentas com esta capacidade por essas gerarem vantagens competitivas sustentáveis.

O processo de extração de conhecimento não explícito a partir de bancos de dados é chamado *Knowledge Discovery in Databases* (KDD). A fase mais

importante desse processo é a mineração de dados ou *Data Mining*, responsável por extrair, interpretar e relacionar informações provenientes de bancos de dados. Entre as diversas técnicas sendo estudas e aplicadas, as relacionadas à linha de pesquisa denominada inteligência computacional vem obtendo excelentes resultados. Algoritmos genéticos, redes neurais artificiais e sistemas fuzzy vêm sendo intensamente investigados, sendo que a combinação dessas técnicas em sistemas híbridos tem chamado mais atenção recentemente.

Sistemas neuro-fuzzy que aliam a capacidade de aprendizado das redes neurais artificiais com a facilidade de interpretação lingüística dos sistemas fuzzy estão sendo amplamente utilizados, especialmente em tarefas de classificação de padrões e previsão de séries temporais. Isto porque esses sistemas possuem características importantes, entre elas: a aplicabilidade dos algoritmos de aprendizado utilizados pelas Redes Neurais; a possibilidade de realizar a integração de conhecimentos implícito e explícito; e facilita a interpretação e o entendimento do mesmo, principalmente pelo fato da extração do conhecimento se dar através de regras fuzzy.

Após a análise de sistemas com essas características, o sistema Neuro-Fuzzy Hierárquico BSP (NFHB) [SOUZ99], destacou-se por ser flexível com relação ao número de variáveis de entrada, além de conseguir gerar a sua própria estrutura automaticamente. Entretanto, esse sistema possui uma grande quantidade de parâmetros de configuração, fato este que implica na necessidade de um especialista para a realização da parametrização adequada. Sendo assim, o usuário final do sistema neuro-fuzzy fica impedido, na maioria das vezes, de utilizá-lo por ser tecnicamente leigo.

1.2. Objetivos

Os principais objetivos deste trabalho são, portanto, modelar e desenvolver um sistema híbrido neuro-fuzzy-genético para a realização automática de tarefas de mineração de dados, ou seja, um sistema que seja transparente para o usuário final no momento em que a mineração de dados esteja sendo realizada, já que toda a parametrização dos modelos híbridos utilizados já estará sendo realizada automaticamente. Além disso, outros objetivos deste trabalho foram:

- Avaliar técnicas de seleção de atributos de entrada em sistemas de mineração de dados;

- Avaliar o sistema neuro-fuzzy-genético em problemas de classificação de padrões e previsão de séries temporais, utilizando dados reais e *benchmarks*.

1.3. Descrição do Trabalho

Este trabalho teve como principais etapas: estudo das técnicas atuais de mineração de dados, identificando, entre essas, quais apresentam melhores resultados; estudo do sistema NFHB original, visando identificar seus principais parâmetros; definição e implementação do modelo neuro-fuzzy-genético; por fim, estudo de casos.

No estudo das técnicas atuais de mineração de dados, foram avaliados métodos estatísticos como Classificadores Bayesianos [SHEN93][CHEE96] e Redes Bayesianas [HECK96]. Também foram estudadas as árvores de decisão, onde a idéia principal é dividir o problema em outros menores para atingir o objetivo principal. Dentre as árvores de decisão estudadas, são evidenciados os modelos: ID3 [QUIN87] e suas evoluções como ID4 e C4.5 [QUIN93a][QUIN93b], CART (*Classification and Regression Trees*) [BREI84] e FID3.1 [JANI99] o qual representa árvores de decisão fuzzy. Outra abordagem estudada é a Associação de Regras [KARU97] que busca encontrar regras de implicação do tipo "X implica em Y", a partir de transações de bancos de dados. Técnicas de Algoritmos Genéticos também são estudadas, entre elas o *Rule Evolver* e o *GA-Miner* [FLOC95]. Ainda na área de inteligência computacional são avaliadas as redes neurais artificiais [ANDR96][NEUM98], os sistemas fuzzy [ZADE65] e, por fim, os sistemas híbridos neuro-fuzzy [JANG93][JANG95][KRUS95], destacando o sistema NFHB [SOUZ99].

O sistema NFHB apresenta importantes características além das comuns a todos os sistemas neuro-fuzzy, como flexibilidade quanto ao número de entradas e capacidade de criar a sua estrutura automaticamente. Deste modo, esse sistema foi escolhido como base para a criação de um sistema automático de mineração de dados. O NFHB foi estudado mais profundamente para avaliação de seu funcionamento, objetivando definir a melhor maneira de otimizá-lo sem a intervenção de um especialista. Dentre as características estudadas foram avaliados: o particionamento BSP, a célula básica NFHB, a arquitetura NFHB e os métodos de atualização dos pesos fuzzy utilizados (Mínimos Quadrados Ordinários e RProp) [SOUZ99]. A partir deste estudo, foram obtidos os

parâmetros que deveriam ser otimizados e, conseqüentemente, configurados automaticamente no sistema, além de identificar a viabilidade computacional de tal otimização.

A seleção das variáveis a serem utilizadas como entrada para o sistema NFHB é um fator essencial para o sucesso dos resultados. Quais variáveis devem ser utilizadas e, principalmente, a ordem em que devem ser fornecidas ao sistema são fatores que podem modificar substancialmente os resultados finais. Deste modo, métodos para realizar estas tarefas foram avaliados, entre eles: o método baseado no modelo Anfis [SOUZ99], utilizado no NFHB original; o método LSE (*Least Square Error*) [CHUN00], onde se busca ordenar as variáveis de acordo com a sua contribuição para a saída; e o método SIE (*Single-Input Effectiveness*) [CAO 97], onde as variáveis não só são ordenadas mas também podem ser eliminadas caso sejam desprezíveis para o sistema. A escolha do melhor método de seleção de variáveis para cada conjunto de dados foi considerada como mais um parâmetro a ser otimizado.

Na coevolução genética vários algoritmos genéticos evoluem em conjunto, normalmente em paralelo, buscando resolver problemas menores mais eficientemente, de forma a gerar um melhor resultado final para o problema principal [PAGI02]. Como o sistema NFHB possui muitos parâmetros distintos, a coevolução surgiu como um método capaz de realizar com excelência a complexa tarefa de otimização dos mesmos. Várias abordagens de coevolução foram analisadas [POTT00][DELG02][HERR99][DELG02][JENS01][STAN02], e dentre elas, a coevolução hierárquica [DELG02] destacou-se como a mais adequada para o problema em questão.

Um modelo neuro-fuzzy-genético baseado em coevolução hierárquica foi criado para a melhor otimização dos parâmetros do sistema NFHB. Quatro algoritmos genéticos (AGs) foram modelados para otimizar diferentes parâmetros, porém com alguma relação entre si. Foi criada uma hierarquia de AGs e, a partir desta arquitetura, o fluxo de coevolução e de troca de informações entre os mesmos foi definido. Também nesta etapa foram definidas as funções de avaliação, os operadores genéticos utilizados e a configuração dos parâmetros dos AGs (GAP, probabilidade de *crossover*, mutação, entre outros).

A etapa final consistiu na realização de estudos de casos, onde tanto problemas de classificação de padrões como previsão de séries temporais foram abordados. Para a tarefa de previsão, foram utilizadas séries de carga elétrica reais, obtidas das empresas Cerj, Cemig, Light, Furnas, Copel e Eletropaulo.

Nesses casos, o sistema automático foi avaliado segundo os valores calculados de RMSE (*Root Mean Square Error*), MAPE (*Mean Absolute Percentage Error*) e Utheil. Para a classificação de padrões, foram utilizadas bases do tipo *benchmark* como *Glass Data*, *Wine Data, Pima Indian Diabetes* e *Bupa Liver Disorders*, sendo o sistema neuro-fuzzy-genético avaliado segundo o MAPE, o RMSE e a porcentagem de acerto na classificação. A partir dos resultados obtidos foram realizadas comparações com os resultados do sistema NFHB original configurado por um especialista [SOUZ01]. Essa comparação foi realizada para a certificação e o encorajamento do uso do sistema proposto. Conforme será visto no capítulo 6, os resultados obtidos demonstram que o sistema automático apresenta resultados semelhantes (em muitos casos melhores) aos obtidos por um especialista, portanto, atingindo o objetivo de unir eficácia com automação.

1.4. Organização da Tese

Esta dissertação possui mais seis capítulos conforme descrito a seguir.

O capítulo 2 apresenta um estudo das técnicas de mineração de dados existentes.

O capítulo 3 descreve os aspectos mais importantes do sistema NFHB e define os principais parâmetros do sistema. Ainda neste capítulo são descritos os métodos de seleção de variáveis utilizados.

O capítulo 4 apresenta os modelos coevolutivos tradicional e hierárquico, detalhando sua características e possibilidades.

O capítulo 5 introduz o sistema híbrido neuro-fuzzy-genético proposto para a mineração automática de dados. São descritos o modelo de coevolução criado, sua arquitetura, funcionamento e configuração.

O capítulo 6 apresenta os estudos de casos realizados, descrevendo os dados utilizados, os resultados obtidos para classificação e previsão e as comparações realizadas com resultados obtidos anteriormente.

No capítulo 7 são apresentadas as conclusões do trabalho e propostos estudos futuros sobre o tema.