1
Introdução

1.1.
Motivação

O cérebro possui incrível capacidade de processamento. Sua estrutura altamente complexa, não linear e paralela, permite executar alguns tipos de tarefas mais rapidamente do que qualquer computador já construído, embora seja superado pelas máquinas atuais nas operações matemáticas ou repetitivas. Para exemplificar a incrível capacidade do cérebro humano, o sistema visual precisa criar uma representação do ambiente e extrair toda a informação necessária para interação com o mesmo em uma fração de segundo. Esse exemplo está entre as tarefas rotineiras de reconhecimento perceptivo do cérebro humano, que são feitas em pouco mais de 100ms, enquanto que um computador pode levar horas, ou até dias (HAYKIN, 2007).

As redes neurais artificiais têm suas bases em várias disciplinas: neurociência, estatística, ciência da computação, engenharia, etc., e tentam imitar a capacidade de reconhecimento do cérebro biológico (HAYKIN, 2007). Elas são utilizadas em muitas áreas, como modelagem e análise de séries temporais, reconhecimento, agrupamento e classificação de padrões, processamento de sinais e controle, mas, assim como o cérebro biológico, precisam ser treinadas e isso pode consumir muito tempo.

Com o aumento da quantidade de informação e a menor disponibilidade de tempo, os sistemas baseados em redes neurais precisam responder cada vez mais rapidamente com cada vez mais informação. Mais do que isso, elas devem ser extremamente flexíveis de forma a se adaptarem aos mais diferentes tipos de problemas, minimizando os custos com modelagem, treinamento, desenvolvimento e implantação.

Algumas bibliotecas de redes neurais já foram desenvolvidas, como a do Matlab ou do Scilab, que são configuradas por meio de *scripts* e, com isso, oferecem muita flexibilidade. Entretanto, esse tipo de configuração dificulta a modela-

PUC-Rio - Certificação Digital Nº 0812723/CA

gem dos problemas, uma vez que o arquiteto da solução precisa ter conhecimento específico do *script* daquele programa. Além disso, o uso de *script* prejudica muito o desempenho, uma vez que este precisa ser interpretado, compilado e, só então, processado.

Por outro lado, existem programas que possuem um ambiente gráfico, como é o caso do Weka. Entretanto, esses programas, normalmente, oferecem pouca ou nenhuma flexibilidade com relação à quantidade de algoritmos de treinamento e modelos de redes neurais. O Matlab possui tanto um ambiente gráfico (através do NNtool) quanto um ambiente de programação por *script*, conforme já foi dito anteriormente, mas seu desempenho deixa muito a desejar. Os trabalhos envolvendo CUDA e redes neurais estão em fase inicial de desenvolvimento e apresentam apenas treinamentos com gradiente decrescente simples, como por exemplo (LAHABAR, AGRAWAL e NARAYANAN, 2009), e outros casos apresentam apenas a propagação de redes, como por exemplo, (JANG, PARK e JUNG, 2008), mas sem possibilidade de configuração de modelos ou de treinamentos.

PUC-Rio - Certificação Digital Nº 0812723/CA

1.2. Objetivos

A presente dissertação propõe uma biblioteca de componentes de redes neurais com dois principais objetivos: flexibilidade e desempenho. O primeiro objetivo requer que a biblioteca seja completamente configurável e suporte situações não previstas como uma topologia nova, na qual os neurônios possuem outro tipo de comportamento ou suas camadas sejam organizadas de forma não convencional. O segundo objetivo é o mais desafiador, pois exige que a biblioteca forneça suporte ao treinamento de sistemas com várias redes ou com vários comitês de redes e utilize bases de dados muito grandes em tempos muito curtos. Essa restrição de desempenho implica no desenvolvimento de um algoritmo de treinamento que rode em placas gráficas, de modo a paralelizar ao máximo a execução do problema.

Além disso, o trabalho apresenta um sistema de janelas gráficas que facilita a modelagem, o treinamento e a implantação das redes neurais, tornando estes procedimentos mais rápidos e disponíveis para diversos tipos de usuário.

1.3. Descrição do trabalho

Este trabalho se desenvolveu de acordo com os seguintes passos: (a) estudo dos modelos de redes neurais existentes e seus processos de treinamento com os diferentes tipos de algoritmos; (b) levantamento das vantagens e desvantagens dos diversos sistemas existentes no mercado, e estudo das áreas de engenharia de software e de computação de alto desempenho; (c) definição e desenvolvimento de uma biblioteca orientada a objetos com capacidade de modelar os diversos tipos de redes neurais e de treinamentos, priorizando o desempenho; (d) desenvolvimento de uma ferramenta gráfica baseada na biblioteca de redes definida e avaliação dos resultados por meio de um estudo de caso com sete experimentos.

O estudo sobre as redes neurais artificiais constituiu-se do levantamento de material bibliográfico sobre os diversos modelos de redes neurais existentes, suas aplicações e os principais algoritmos de treinamento.

O levantamento dos sistemas existentes exigiu uma busca de dois principais grupos de software: os gratuitos e os pagos. Percebeu-se que existe uma grande carência de programas gratuitos que possam ser utilizados em projetos na área de apoio à decisão. Além disso, a grande maioria dos sistemas pagos, baseados em redes neurais, apresenta uma arquitetura pobre, com relação à interface com o usuário e configurações disponíveis, e não respondem rapidamente em vários problemas reais.

O desenvolvimento de uma biblioteca, que respeitasse as restrições de arquitetura e desempenho, exigiu o estudo de duas áreas da computação: engenharia de software e computação de alto desempenho. Os conceitos da primeira foram usados na modelagem, arquitetura e ciclo de vida do sistema e os conceitos da segunda serviram para modificar partes chaves do código visando à redução do tempo de processamento dos algoritmos de treinamento.

Outro ponto importante percebido nos sistemas de redes neurais existentes foi a interface. Essa, normalmente, exige noções de programação dos clientes ou possui uma interface estática que facilita a utilização do sistema, mas torna a personalização dos modelos limitada ou impossível.

Finalmente, foram realizados sete experimentos de um estudo de caso para se avaliar o desempenho do sistema desenvolvido neste trabalho. Foram obtidos

PUC-Rio - Certificação Digital Nº 0812723/CA

seus gráficos comparativos e o desempenho das soluções foi comparado com os resultados das diferentes abordagens presentes na bibliografia.

1.4. Estrutura da Dissertação

Esta dissertação está dividida em cinco capítulos adicionais, descritos a seguir.

O Capítulo 2 apresenta os principais conceitos da Engenharia de Software, incluindo o levantamento de requisitos, projeto e arquitetura, os padrões de projeto utilizados no trabalho e o arcabouço da Microsoft, que foi utilizado como base para o desenvolvimento deste trabalho.

O Capítulo 3 descreve as várias abordagens utilizadas na história da computação para se melhorar o desempenho dos sistemas e como cada uma funciona.

O Capítulo 4 apresenta a biblioteca de componentes proposta nesta dissertação, detalhando a sua estrutura, seus componentes de redes neurais e algoritmos de treinamento, além de uma ferramenta gráfica para criação de soluções com redes neurais artificiais. Por último, são apresentadas as modificações feitas na arquitetura da biblioteca de modo que esta fosse executada, em parte, por uma placa gráfica.

O Capítulo 5 descreve os experimentos do estudo de caso com seus respectivos modelos de redes neurais e discute os resultados obtidos em função das referências bibliográficas.

Por fim, o Capítulo 6 apresenta as conclusões do trabalho e sugere novas direções possíveis para a continuação da pesquisa apresentada.

PUC-Rio - Certificação Digital Nº 0812723/CA