MODELOS DE APRENDIZAGEM DE MÁQUINA NA CLASSIFICAÇÃO DE CARACTERES MANUSCRITOS

Cheila Maria Bergamini(1); Priscila Vriesman Araujo(1) & Giovani Motter(2)

(1) Mestrandas pelo Programa de Pós-Graduação em Informática Aplicada da PUC-PR. (2) Professor do cursos de Sistemas de Informação da UNISEP e Ciência da Computação da UNIPAN. Mestrando pela FEEC/UNICAMP.

cheilamaria@gmail.com; priscila.vriesman@gmail.com; gmotter@unisep.edu.br;

Resumo - Atualmente, algoritmos de aprendizagem de máquina são utilizados para auxiliar no desenvolvimento de soluções de inúmeros problemas. Este artigo realiza um estudo sobre os principais modelos e técnicas de aprendizagem para classificação e reconhecimento de caracteres manuscritos, descrevendo as características e aplicações de cada um deles, e demonstrando-as através de experimentos práticos.

Palavras-Chave – Reconhecimento de Padrões, Aprendizagem de Máquina, Redes Neurais, Árvores de Decisão.

MODELOS DE APRENDIZAGEM DE MÁQUINA NA CLASSIFICAÇÃO DE CARACTERES MANUSCRITOS

1. INTRODUÇÃO

Inúmeros são os problemas que são efetuados de forma repetitiva e manualmente, através da ação humana. De forma a automatizar essas atividades, é possível realizar um estudo e posterior desenvolvimento de rotinas computacionais que auxiliem, ou até mesmo substituam o trabalho humano em tais atividades.

A Inteligência Artificial (IA) é uma área da Ciência da Computação responsável pelo estudo e desenvolvimento de tais rotinas computacionais, envolvendo pesquisas relacionadas à aprendizagem e evolução computacional, sistemas especialistas, sistemas de suporte à tomada de decisão, agentes computacionais inteligentes, entre outras áreas.

Desta forma, problemas relacionados ao desenvolvimento de “seres computacionais inteligentes”, sistemas especialistas em uma determinada atividade, sistemas para reconhecimento de padrões, como reconhecimento de imagens, por exemplo, e sistemas de aprendizagem, que são capazes de aprender com suas experiências ao longo do tempo, podem ser solucionados através de uma solução proposta pelas técnicas de IA (FERNANDES, 2005).

O problema de reconhecimento de padrões se aplica em inúmeras situações, que vão desde uma simples identificação de formas geométricas, à identificação de caracteres manuscritos e, até mesmo, de sons e imagens. Baseado no problema de identificação, classificação e reconhecimento de caracteres manuscritos, este trabalho de pesquisa tem como objetivo estudar diferentes formas e técnicas para realizar a validação de caracteres manuscritos, utilizando-se de uma base de treinamento com diversas características, possibilitando uma comparação entre

as diferentes técnicas, identificando a mais apropriada para a classificação de caracteres manuscritos.

O presente trabalho se propõe, portanto, a apresentar uma descrição mais detalhada sobre o problema do reconhecimento de caracteres manuscritos, além de uma breve revisão sobre as técnicas de IA que podem ser utilizadas para solucionar este problema. Na seqüência, com a implementação destas técnicas, serão realizados testes sobre um conjunto de caracteres de treinamento, para posterior análise e conclusão. Finalizando, serão apresentados os resultados obtidos, as conclusões e considerações finais.

2. O PROBLEMA DE RECONHECIMENTO DE CARACTERES MANUSCRITOS

Devido à natureza discreta dos sistemas de computação, o reconhecimento e classificação de caracteres manuscritos é uma tarefa árdua e minuciosa, uma vez que tais caracteres podem apresentar diversas formas e similaridades.

O reconhecimento ou classificação de um determinado caractere manuscrito pode ser realizado através da verificação de suas características, e posterior comparação a um conjunto de amostras válidas e reconhecidas.

O NIST (*National Institute of Standards and Technology*) armazena em uma base de dados especial (*Special Database* 19 – SD19) um amplo conjunto de caracteres e formas manuscritas para realização de reconhecimento e treinamento (NIST, 2006). Os caracteres neste banco de dados diferenciam-se uns dos outros através de um conjunto de características e do caractere correspondente.

Este trabalho de pesquisa se justifica pela necessidade de estudo das diferentes formas e técnicas para realizar a classificação e reconhecimento de objetos, mais especificamente de caracteres manuscritos, baseando-se na análise de características das amostras.

Para realização deste trabalho, utilizou-se um conjunto de caracteres manuscritos reais, maiúsculos e minúsculos, extraídos

da base de dados do NIST SD19, identificando, para cada elemento deste conjunto, 108 características distintas.

3. ALGORITMOS DE APRENDIZAGEM

Inúmeros são os conceitos de aprendizagem que podemos encontrar na literatura. No entanto, todos convergem para uma mesma conclusão. Mitchell (1997) apresenta o conceito de aprendizagem computacional como sendo "uma forma ou técnica para realizar um treinamento sobre um determinado conjunto de treinamento", ou ainda, como "uma solução algorítmica que permite o aprendizado com o passar do tempo" possibilitando a melhoria dos resultados com as experiências obtidas.

Para Russel e Norvig (2004), o conceito de aprendizagem de software está relacionado à experiência obtida ao longo do tempo. Ainda, os autores descrevem que a "aprendizagem pode variar desde a memorização trivial da experiência até a criação de novas teorias".

A aprendizagem de máquina pode ser vista como uma ferramenta poderosa, porém não existe somente um algoritmo que apresente o melhor desempenho para todos os problemas. São vários algoritmos estudados em IA, e cada um apresenta suas particularidades, pontos fortes e fracos, relacionados ao desempenho e limitações.

Na seqüência, serão apresentados, de uma forma breve, os principais algoritmos e métodos de aprendizagem, os quais serão utilizados nas experiências práticas deste trabalho: aprendizagem baseada em instâncias; aprendizagem em árvore de decisão; redes neurais.

3.1. Aprendizagem Baseada em Instâncias

Para Russel e Norvig (2004), os métodos de aprendizagem baseada em instância permitem que a complexidade de hipóteses cresça com os dados, ou seja, quanto mais dados existirem, mais complicada a hipótese pode ser.

Os métodos de aprendizagem baseada em instâncias (ou aprendizagem baseada na memória) elaboram hipóteses

diretamente a partir das próprias instâncias de treinamento, e são classificados em dois modelos distintos: o modelo do vizinho mais próximo e o modelo de núcleo. Devido à característica do modelo do vizinho mais próximo realizar a consulta do elemento do conjunto de treinamento que mais se aproxima ao elemento analisado para classificação, este trabalho realizará um estudo mais detalhado sobre este modelo, apenas.

A técnica do vizinho mais próximo é simples, e para compreendê-la não é necessário muitos cálculos. O primeiro passo é a identificação das características essenciais para resolução do problema, de tal modo que a medida de distância entre o novo caso e os casos existentes na base de dados possam ser medidas. A classificação é feita conforme a distância calculada entre o novo elemento e todos os elementos existentes no conjunto de treinamento (FERNANDES, 2005).

Segundo Russel e Norvig (2004), a idéia-chave de modelos de vizinho mais próximo é que as propriedades de qualquer ponto da entrada x específico têm probabilidade de serem semelhantes às propriedades de pontos na vizinhança de x.

A similaridade entre o novo elemento e um elemento existente é determinada para cada atributo. Esta medida deve ser obtida através da somatória das distâncias euclidianas entre o par de atributos do novo elemento e o elemento existente no conjunto de treinamento, para todos os atributos existentes. Este cálculo de distância deve ser efetuado entre o novo elemento e todos os elementos existentes no conjunto de treinamento; aquele elemento do conjunto de treinamento que apresentar a menor distância em relação ao novo elemento é chamado de vizinho mais próximo.

Portanto, este cálculo deve ser repetido para toda a base de dados, para a obtenção de seu “ranking”, identificando vizinhos mais próximos, ou seja, aqueles que apresentam as características mais próximas.

Conforme Fernandes (2005), a técnica do vizinho mais próximo talvez seja a mais usada para estabelecimento de similaridade, pois a grande maioria das ferramentas disponível a utiliza. É a técnica mais indicada para problemas que possuam uma base de casos pequena e poucos atributos indexados, por

causa do volume de cálculos necessários para a determinação de cada atributo indexado a cada um dos casos.

O algoritmo k-NN (k *Nearest Neighbor*, ou k vizinhos mais próximos) é uma evolução do modelo do vizinho mais próximo, onde se torna possível a realização de uma votação entre os k vizinhos mais próximos a um elemento, identificando sua aceitação ou não, isto é, identificando se a maioria dos vizinhos mais próximos apresentam a mesma função alvo que o novo elemento.

Resumindo, a aprendizagem baseada em instância consiste em armazenar as instâncias de treinamento, calcular a distância entre a instância de treinamento e a instância desconhecida, avaliando o valor da função a partir das instâncias mais próximas.

3.2. Aprendizagem em Árvores de Decisão

A indução de árvores de decisão, segundo Russel e Norvig (2004), é uma das formas mais simples, e ainda assim bem-sucedidas, de algoritmos de aprendizagem.

A árvore de decisão é usada para criar regras com os nós, servindo como pontos de decisão. Assim, a árvore de decisão alcança sua decisão executando uma seqüência de testes, onde cada nó interno na árvore corresponde a um teste do valor de uma das propriedades, e as ramificações a partir do nó são identificadas com os valores possíveis. Cada nó da folha na árvore especifica o valor a ser retornado se aquela folha for alcançada.

As árvores de decisão são completamente expressivas dentro da classe de linguagens proposicionais, isto é, qualquer função booleana pode ser escrita como uma árvore de decisão. Porém, para representar o conteúdo de uma tabela-verdade em uma árvore de decisão seria necessária uma árvore exponencialmente grande, uma vez que as tabelas-verdade têm exponencialmente muitas linhas (RUSSEL & NORVIG, 2004).

3.3. Redes Neurais Artificiais (RNA)

As RNAs são inspiradas na estrutura do cérebro humano, com o objetivo de apresentar características similares a eles: aprendizado, associação, generalização e abstração. Elas são compostas por diversos elementos, os processadores (neurônios

artificiais), altamente interconectados, que efetuam um número pequeno de operações simples e transmitem seus resultados aos processadores vizinhos (RUSSEL & NORVIG, 2004).

A arquitetura das redes neurais é tipicamente organizada em camadas (FERNANDES, 2005). Segundo Russel e Norvig (2004), as Redes Neurais são classificadas quanto ao número de camadas: as redes neurais de alimentação direta de uma única camada, ou redes *perceptron*, e as redes neurais de alimentação direta de várias camadas, ou redes *perceptron* multi-camadas (MLP – *Multi-Layer Perceptron*).

O modelo de RNA mais utilizado atualmente é o MLP (FERNANDES, 2005), treinada com o algoritmo *backpropagation*.

O treinamento das redes MLP com *backpropagation* pode demandar muitos passos no conjunto de treinamento, resultando um tempo de treinamento consideravelmente longo. Segundo Fernandes (2005), o algoritmo *backpropagation momentum* apresenta uma pequena modificação em relação ao algoritmo *backpropagation* de maneira de aumentar a taxa de aprendizado.

4. EXPERIMENTOS

Como visto, diversas são as técnicas e modelos existentes para classificação e reconhecimento. Este trabalho de pesquisa apresenta como objetivo um estudo sobre as principais técnicas utilizadas para classificação e reconhecimento de padrões, identificando suas características e aplicações, e realizando experimentos práticos sobre o problema de reconhecimento e classificação de caracteres manuscritos.

Esta seção realizará uma breve descrição de como foram realizados os experimentos práticos, utilizando algoritmos de aprendizagem baseados em instâncias, em árvores de decisão, e redes neurais, além da apresentação dos resultados obtidos, considerações e conclusões.

Para realização dos experimentos, utilizou-se uma base de dados para treinamento de caracteres manuscritos SD19 fornecida pelo NIST (2006). Esta base de dados serviu como modelo para a classificação de elementos de outras bases de

dados, contendo elementos distintos: uma base de validação, e outra de testes.

A tabela 1 apresenta algumas informações resumidas sobre a estrutura dos bancos de dados utilizados nos experimentos, sendo que todas armazenam informações sobre 52 tipos distintos de caracteres manuscritos maiúsculos e minúsculos:

Tabela 1: Estruturas dos bancos de dados utilizados nos experimentos.

Base de dados / Arquivo	Nº de Amostras	Nº de Características
Treinamento	74.880	108
Validação	23.670	108
Testes	23.941	108

4.1. Algoritmo do vizinho mais próximo k-NN

Para a realização do experimento utilizando o algoritmo do vizinho mais próximo k-NN no modelo baseado em instâncias, foi necessária a implementação do algoritmo. De forma a apresentar uma melhor performance na leitura dos arquivos de dados, e também nos procedimentos e cálculos matemáticos previstos pelo algoritmo, optou-se pela implementação utilizando a linguagem de programação ANSI C.

O programa implementado apresenta módulos para a realização da leitura dos arquivos, cálculo das distâncias entre os elementos, além da identificação os vizinhos mais próximos e o cálculo da “votação da maioria”. Os parâmetros de entrada necessários para execução do programa devem ser informados através de constantes declaradas no código do programa.

4.2. Aprendizagem de Árvores de Decisão

Para a realização do experimento sobre o modelo de aprendizagem de árvores de decisão, utilizou-se o software C4.5. O C4.5 é um software de classificação utilizado para a representação de árvores de decisão, e tem a função de construir modelos de classificação indutivamente a partir dos dados analisados por meio de generalização de exemplos específicos (BERNARDES, 2001).

Possuindo versões para plataformas Windows e Linux, o C4.5 é muito utilizado em pesquisas acadêmicas para auxiliar nos problemas de classificação e criação de árvores de decisão.

Porém, sua versão para Windows apresenta limitações, as quais não são encontradas na versão para Linux.

Para se aplicar o processo C4.5 no experimento, foram necessários efetuar os seguintes procedimentos, em uma estação Sun multiprocessada com 4Gb de memória RAM, e Sistema Operacional Linux Fedora Core 4.

4.3. Redes Neurais

No experimento realizado para redes neurais, foi utilizado para treinamento, inicialmente, o simulador JavaNNS multi-plataforma, porém os resultados obtidos com este simulador não foram satisfatórios, uma vez que os resultados esperados estariam na faixa entre 0 e 1, e com a utilização deste simulador os resultados foram sempre maiores do que 1. Não se sabe a razão pela qual o simulador apresentou tais resultados, visto que os mesmos parâmetros utilizados para o treinamento neste simulador também foram utilizados no simulador SNNS, um simulador para plataforma Linux, que apresentou resultados satisfatórios e performance superior comparando-se ao primeiro.

O modelo de rede neural utilizado no experimento foi o Perceptron multi-camadas, com algoritmo de treinamento *backpropagation momentum*. Este algoritmo é indicado para treinamento de grandes conjuntos de dados, como citado na seção anterior.

Para inicializar a arquitetura da rede, estabeleceu-se uma rede composta por três camadas: entrada, saída e camada oculta. A camada de entrada é formada por 108 neurônios, representando as 108 características existentes para cada elemento do conjunto de treinamento; a camada de saída é formada por 52 neurônios, referindo-se ao conceito alvo, ou seja, correspondendo aos 52 possíveis caracteres maiúsculos e minúsculos; e a camada oculta, para armazenamento dos pesos intermediários, composta por 80 neurônios, um valor médio entre o número de neurônios das camadas de entrada e de saída.

A figura 1 ilustra o gráfico gerado pelo SNNS para classificação do conjunto de treinamento, utilizando uma rede neural com os parâmetros citados acima, em uma simulação caracterizada por 400 ciclos.

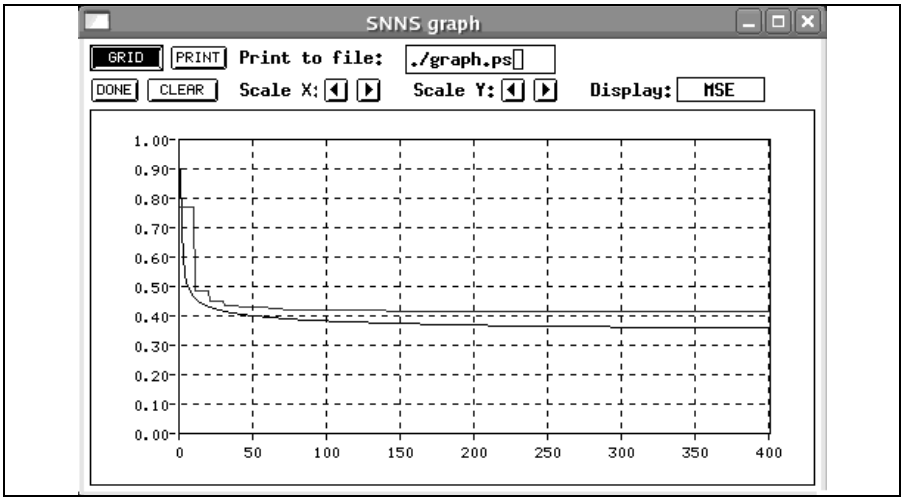

Figura 1: Evolução do treinamento utilizando RNA.

5. CONCLUSÃO E CONSIDERAÇÕES FINAIS

Realizando uma comparação com os resultados obtidos entre os experimentos realizados com os diferentes métodos e técnicas de aprendizagem, podemos relacionar os dados conforme descreve a tabela 2.

Tabela 2: Resultados obtidos através dos experimentos realizados.

Modelo	Taxa de Acertos	Taxa de Erros
Aprendizagem Baseada em Instâncias	68,24 %	31,76 %
Aprendizagem de Árvore de Decisão	53,3 %	46,7 %
Rede Neural	74,23 %	25,77 %

Como visto na tabela acima, os resultados obtidos nos experimentos foram bem sucedidos, porém o modelo de rede neural apresentou os melhores resultados comparados aos demais modelos estudados neste trabalho de pesquisa. Isto se justifica, uma vez que um Perceptron Multi-Camadas possibilita a identificação do ponto de convergência entre a generalização e a especialização, dependendo do conjunto de dados de treinamento. Este ponto de convergência apresenta uma aproximação do melhor resultado.

O modelo de aprendizagem baseada em instâncias k-NN também apresentou resultados satisfatórios para o problema de classificação de caracteres manuscritos em um banco de dados grande. Quanto mais próximos do elemento, maior é a probabilidade de acerto na classificação e reconhecimento.

Entretanto, o modelo de aprendizagem baseado em árvores de decisão não apresentou bons resultados, uma vez que para este exemplo de aplicação utiliza um conjunto de treinamento exponencialmente grande, é necessário a criação de uma árvore também exponencialmente grande. Devido ao tamanho da árvore, este modelo generaliza muito, e não apresenta uma boa precisão para classificação dos elementos.

6. REFERÊNCIAS

BERNARDES, R. M.. C4.5: Um Recurso para Geração de Árvores de Decisão. Instruções Técnicas. Ministério da Agricultura, Pecuária e Abastecimento, 2001.

FERNANDES, A. M. R. Inteligência Artificial: Noções Gerais. Florianópolis : Visual Books, 2005.

MITCHELL, TOM M. Machine Learning. McGraw-Hill, 1997.

NIST – NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. NIST Handprinted Forms and Characters Database: Special Database 19. Disponível em: <http://www.nist.gov/srd/nistsd19.htm> . Acesso em: 30 abr. 2006.

OLIVEIRA, R. S. Sistemas Inteligentes – Fundamentos e Aplicações. Barueri, SP: Manole, 2003, pp. 123–135.

RUSSEL, S.; NORVIG, P. Inteligência Artificial. Rio de Janeiro : Campus, 2004.