

# Uso de Técnicas de Aprendizado de Máquina no Reconhecimento de Dígitos Manuscritos em Laudos Médicos

Adewole Caetano<sup>1</sup>, Ana C. Lorena<sup>1</sup>, Newton Spolaôr<sup>2</sup>, Willian Zalewski<sup>2</sup>, Huei D. Lee<sup>2</sup>, André G. Maletzke<sup>3</sup>, Wu Feng Chung<sup>2</sup>, Cláudio Saddy<sup>4</sup>, João J. Fagundes<sup>4</sup>, Rodrigues Coy<sup>4</sup>

<sup>1</sup> Centro de Matemática, Computação e Cognição – CMCC, Universidade Federal do ABC – UFABC, Santo André – SP, Brasil

<sup>2</sup> Centro de Engenharia e Ciências Exatas – CECE, Universidade Estadual do Oeste do Paraná – Unioeste, Laboratório de Bioinformática – LABI, Parque Tecnológico Itaipu – PTI – Foz do Iguaçu - PR, Brasil

<sup>3</sup> Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo – USP, Laboratório de Inteligência Computacional – LABIC, São Carlos - SP, Brasil

<sup>4</sup> Faculdade de Ciências Médicas – FCM, Universidade Estadual de Campinas – Unicamp, Campinas, SP, Brasil

{adewole.caetano, [ana.lorena@ufabc.edu.br](mailto:ana.lorena@ufabc.edu.br),  
newtonspolaor@gmail.com, [hueidianalee@gmail.com](mailto:hueidianalee@gmail.com)}

**Abstract.** *Medicine has highly benefited by the use of computational techniques, which are often employed in the analysis of data generated in medical clinics. Among the computational techniques used in these analyses are those from Machine Learning (ML). This paper presents the use of k-Nearest Neighbors, Artificial Neural Networks and Support Vector Machines, three ML techniques, in the recognition of handwritten digits in medical reports.*

**Resumo.** *A medicina tem sido muito beneficiada pelo uso de técnicas computacionais, que geralmente são empregadas na análise de dados produzidos em exames médicos. Dentre as técnicas computacionais utilizados nestas análises estão as técnicas de Aprendizado de Máquina (AM). Neste trabalho é apresentada a utilização do algoritmo k-vizinhos mais próximos, de Redes Neurais Artificiais e de Máquinas de Vetores de Suporte, três técnicas de AM, no reconhecimento de dígitos manuscritos em relatórios médicos.*

## 1. Introdução

Atualmente, o desenvolvimento tecnológico possibilitou o surgimento de novos artifícios para melhorar o desempenho dos meios de armazenamento de dados digitais.

Com isso, torna-se necessário o estudo de métodos e técnicas que possam auxiliar na análise de dados por meio de ferramentas computacionais.

Um dos métodos que oferece apoio na análise de dados é o processo de Descoberta de Conhecimento em Bases de Dados (DCBD), o qual pode ser aplicado para construir modelos que expressem o conhecimento contido em grandes volumes de dados, auxiliando especialistas no processo de tomada de decisão.

Entre as diversas áreas em que a aplicação do processo de DCBD pode ser útil está a medicina. Contudo, nessa área os dados encontram-se, freqüentemente, em formatos desestruturados ou semi-estruturados, dificultando a análise por meio de métodos computacionais. Esses dados podem estar representados, por exemplo, em laudos médicos ou formulários impressos contendo diversas informações como sintomatologia e histórico de pacientes. Para que esses dados possam ser analisados por meio da aplicação de métodos computacionais, como o processo de DCBD, é necessário que estejam representados em um formato estruturado, freqüentemente em um formato atributo-valor (Rezende, 2003).

Este trabalho está inserido dentro de um projeto de aplicação de técnicas inteligentes no mapeamento de dados médicos para conjuntos de dados no formato atributo-valor. Uma metodologia proposta para o mapeamento de formulários médicos foi aplicada com sucesso a alguns estudos de caso (Maletzke et al., 2007). No entanto, um dos desafios a ser trabalhado é o tratamento de caracteres numéricos escritos manualmente, o qual se tornou objetivo principal deste trabalho. Neste trabalho, três técnicas inteligentes de Aprendizado de Máquina (AM) (Mitchell, 1997) foram comparadas no reconhecimento de dígitos manuscritos: k-vizinhos mais próximos, Redes Neurais Artificiais e Máquinas de Vetores de Suporte.

O sucesso neste segmento de pesquisa tornará possível a coleta de dados contínuos presentes nos formulários citados, tornando o conjunto de dados mais robusto e completo e, posteriormente, facilitando a aplicação de métodos para a extração de conhecimento dentro do processo de DCBD.

Esse artigo encontra-se estruturado como segue: na Seção 2 o processo de reconhecimento de dígitos manuscritos por meio de técnicas de AM é descrito. Na Seção 3 são apresentados os experimentos realizados neste projeto no intuito de desenvolver um sistema para o reconhecimento de dígitos manuscritos em laudos médicos. Finalizando, uma conclusão é realizada na Seção 4.

## **2. Aprendizado de Máquina no Reconhecimento de Dígitos Manuscritos**

O princípio das técnicas de Aprendizado de Máquina (AM) consiste em obter resultados genéricos a partir de um conjunto particular de dados (Mitchell, 1997). Para tal, as técnicas de AM empregam um princípio de inferência denominado indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de exemplos. O aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não-supervisionado.

O tipo de aprendizado focado neste trabalho é o supervisionado, em que se tem a figura de um professor externo, o qual apresenta o conhecimento do ambiente por

conjuntos de exemplos na forma: entrada, saída desejada. O algoritmo de AM extrai a representação do conhecimento a partir desses exemplos. O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente.

Os sistemas de reconhecimento de caracteres manuscritos baseados em técnicas de AM geralmente envolvem três etapas, detalhadas a seguir (Liu et al., 2004).

Na primeira, denominada fase de pré-processamento dos caracteres, ocorre a eliminação de possíveis ruídos das imagens dos dígitos manuscritos. Além disso, elas são padronizadas, por exemplo, em tamanho.

Na segunda etapa é realizada a construção de atributos sobre o caractere, como o número de intersecção de retas que este contém, momentos estatísticos, entre outros. Essa etapa é denominada fase de extração de características.

Na última fase, de classificação dos caracteres, ocorre a geração de um classificador responsável por determinar a classe a que o dígito pertence (0, 1, 2, 3, 4, 5, 6, 7, 8 ou 9). Várias técnicas de AM podem ser empregadas nessa etapa, como as Redes Neurais Artificiais e as Máquinas de Vetores de Suporte (LeCun et al., 1995; Goltsev & Rachkovskij, 2005), redes Bayesianas (Sung et al., 2006) e lógica fuzzy (Hammandlu & Murthy, 2007).

No caso deste trabalho, foram utilizados o algoritmo k-vizinhos mais próximos, Redes Neurais Artificiais e Máquinas de Vetores de Suporte.

O processo de modelagem realizado na geração de um classificador por uma técnica de AM envolve diversas etapas, as quais são ilustradas de maneira simplificada na Figura 1. Resumidamente, o classificador deve ser ajustado, por meio de alterações em seus parâmetros livres, e avaliado.

O conjunto de exemplos fornecido a um algoritmo de AM é usualmente dividido em dois subconjuntos disjuntos, um para o ajuste do modelo, de treinamento, e outro para a sua avaliação, de teste. O subconjunto de treinamento é utilizado no aprendizado do conceito e o subconjunto de teste é utilizado para medir o grau de efetividade do conceito aprendido na previsão da classe de novos dados. Nesse processo, os dados de teste são apresentados ao classificador e as previsões produzidas pelo modelo são comparadas aos rótulos conhecidos, produzindo assim uma estimativa de acerto e erro do modelo.

A seguir é apresentada uma breve introdução às técnicas de AM utilizadas neste trabalho.

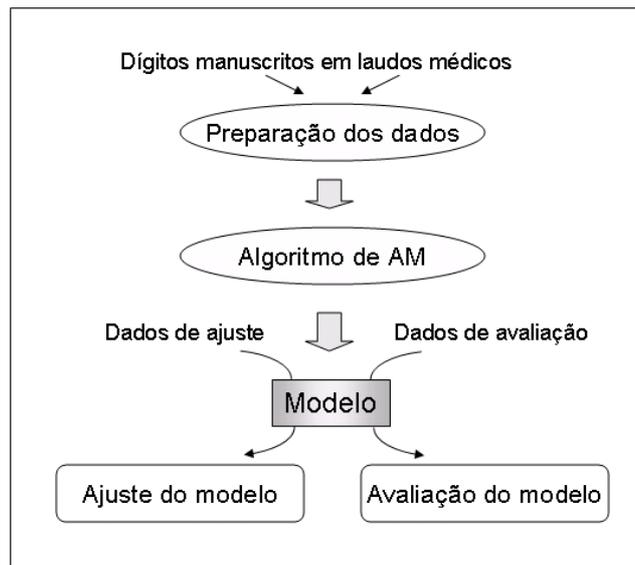


Figura 1 - Esquema simplificado de processo de modelagem por técnica de AM

## 2.1. Máquinas de Vetores de Suporte (*Support Vector Machines*)

As Máquinas de Vetores de Suporte (SVMs – do Inglês *Support Vector Machines*) (Cristianini & Shawe-Taylor, 2000) constituem uma técnica de aprendizado baseada em conceitos da Teoria de Aprendizado Estatístico. Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores ao uso de Redes Neurais Artificiais. Exemplos de sucesso podem facilmente ser encontrados em aplicações diversas, como: categorização de textos, análise de imagens e em Bioinformática.

Em sua forma básica, as SVMs buscam uma fronteira linear capaz de separar os dados de diferentes classes com uma margem máxima de separação. Por meio desse artifício, a SVM possui a característica de possuir boa capacidade de generalização, relacionada à previsão correta da classe de novos dados do mesmo domínio em que o aprendizado ocorreu (Lorena, 2006).

## 2.2. Rede Neural Artificial (Perceptron Multicamadas)

As Redes neurais artificiais (RNAs) (Braga et al., 2000) trabalham no processamento de dados de maneira semelhante ao cérebro humano. Elas são compostas de um sistema de neurônios ou nodos simples ligados por conexões ponderadas. Os nodos computam funções matemáticas sobre as entradas ponderadas que recebem e são divididos em: neurônios de entrada, onde os padrões apresentados à rede; neurônios internos, onde é feita a maior parte do processamento; e neurônios de saída, onde o resultado final é concluído e apresentado. A forma de arranjar neurônios por camadas é chamada de Perceptron Multicamadas e foi projetada para resolver problemas mais complexos que não poderiam ser resolvidos por um único neurônio básico. O comportamento inteligente de uma RNA vem das interações entre as unidades de processamento da rede.

Redes Multicamadas podem usar um grande número de técnicas de aprendizado, sendo que a mais popular é a propagação reversa (*backpropagation*). Neste algoritmo, os valores de saída são comparados com a resposta correta para computar o valor de alguma função-erro pré-definida. Por alguma técnica de correção, o erro é então alimentado de volta na rede. Usando essa informação, o algoritmo ajusta os pesos de cada conexão para reduzir o valor da função erro. Na Figura 2 é ilustrado um exemplo de RNA Perceptron Multicamadas.

As RNAs multicamadas são aproximadoras gerais de funções e são usualmente robustas diante de dados ruidosos.

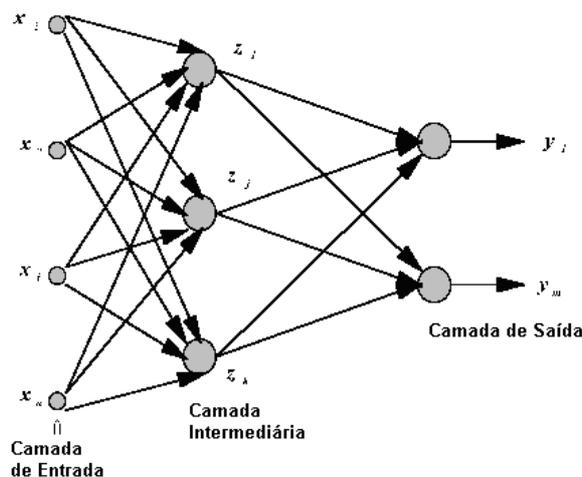


Figura 2 – RNA Perceptron Multicamadas

### 2.3. k-Vizinhos mais próximos (Nearest Neighbor)

Esse algoritmo de classificação baseado no vizinho mais próximo é uma técnica amplamente empregada para reconhecer padrões. O centro de seu funcionamento está em descobrir os vizinhos mais próximos de uma dada instância.

Este classificador pertence a um grupo de técnicas denominadas de *Instance-based Learning*. Ele classifica um dado elemento de acordo com as respectivas classes dos vizinhos mais próximos pertencentes a uma base de treinamentos.

O algoritmo calcula a distância do elemento dado para cada elemento da base de treinamento e então ordena os elementos da base de treinamento do mais próximo ao de maior distância. Dos elementos ordenados selecionam-se apenas os primeiros, que servem de parâmetro para a regra de classificação.

## 3. Experimentos

Para auxiliar no entendimento e desenvolvimento da pesquisa, foram efetuados diversos testes práticos utilizando o software específico para mineração e classificação de dados, WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>), inicialmente com seus parâmetros *default*.

Empregou-se ainda o conjunto de dados MNIST, o qual possui um total de 70000 dígitos, divididos em 60000 para treinamento e 10000 para teste. Deste conjunto de dados, foram extraídas diversas características: *Chaincode Direction*, *Contour Profile*, *End Points*, *Gradient Direction*, *Intersections with Straight Lines*, *Junctions*, *Projection Histogram*, *Radial Codification*, *Zoning*. Detalhes a respeito desses métodos podem ser encontrados em (Zalewski et al. 20007).

Na Tabela 1 são evidenciadas as taxas de classificações corretas para diferentes combinações de métodos de extração de características e classificadores (*Multilayer Perceptron*, *Support Vector Machine* ou *Nearest Neighbor*). A união de todos os métodos de extração de características, denominada *All*, também foi testada.

**Tabela 1 – Resultados das classificações**

EC Method	NN	SVM	MLP
ChainCode Direction	52.65	53.68	<b>58.83</b>
Contour Profile	93.69	94.24	<b>96.04</b>
Gradient Direction	57.90	66.16	66.01
Intersections With Straight Lines	39.22	60.73	<b>66.65</b>
Projection Histogram	87.93	86.18	<b>89.49</b>
Radial Codification	57.17	58.13	<b>59.33</b>
Zoning	<b>87.67</b>	83.63	84.87
End Points	10.43	51.48	<b>54.06</b>
Junctions	23.50	31.06	30.74
All	96.7	98.21	<b>98.67</b>

A análise dos resultados obtidos evidencia o grau satisfatório de desempenho das técnicas de classificação RNAs e SVMs, embora de maneira geral as RNAs tenham se destacado mais fortemente. Observa-se também que nenhum dos métodos de extração de características individual se sobressaiu perante a união de todas as extrações individuais de características.

A RNA obtida com o conjunto *All* foi testada na classificação de 504 dígitos manuscritos em laudos médicos com áreas de escrita previamente delimitadas e obteve resultados considerados satisfatórios de 96.23% de acerto.

Embora existam outros trabalhos que empregam técnicas de AM no reconhecimento de dígitos manuscritos, seus resultados não podem ser diretamente comparáveis aos obtidos neste artigo, devido a diferentes protocolos experimentais seguidos.

#### 4. Conclusão

Neste trabalho foi apresentado o desenvolvimento de um módulo para o reconhecimento de dígitos manuscritos em laudos médicos. Para tal, três técnicas de Aprendizado de Máquina foram comparadas e entre elas as Redes Neurais Artificiais foram escolhidas para a tarefa de classificação final desses dígitos.

Entre os trabalhos futuros advindos dessa pesquisa está a investigação de métodos de seleção de características mais relevantes para a classificação dos dígitos e o

emprego de um processo de ajuste de parâmetros mais refinado para as técnicas de AM utilizadas. Além disso, os dados adquiridos de formulários médicos poderão ser futuramente utilizados para a extração de conhecimento por um processo de DCBD.

## **Agradecimentos**

Os autores gostariam de agradecer ao Programa de Desenvolvimento Tecnológico Avançado (PDTA) da Fundação Parque Tecnológico Itaipu (FPTI) e à Universidade Federal do ABC (UFABC) pelo apoio recebido para a realização deste trabalho.

## **Referências bibliográficas**

Braga, A., Carvalho, A. C. P. L. F., & Ludermir, T. B. (2000). *Redes Neurais Artificiais: Teoria e Aplicações*. Editora LTC.

Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Goltsev, A. & Rachkovskij, D. (2005). Combination of the assembly neural network with a perceptron for recognition of handwritten digits arranged in numeral strings, *Pattern Recognition* (38), No. 3, pp. 315-322.

Hanmandlu, M. & Murthy, O. V. R. (2007). Fuzzy model based recognition of handwritten numerals, *Pattern Recognition* (40), No. 6, pp. 1840-1854.

LeCun, Y. A., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Müller, E., Sackinger, E., Simard, P. Y., and Vapnik, V. N. (1995). Comparison of learning algorithms for handwritten digit recognition. In Foulgeman-Soulié, F. and Gallinari, P., editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'95)*, volume 2, pages 53–60, Nanterre, France.

Liu, C. -L.; Nakashima, K.; Sako, H.; Fujisawa, H. (2004) Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, v. 37. p. 265-279.

Lorena, A. C. (2006). *Investigação de estratégias para a geração de máquinas de vetores de suporte multiclases*. Instituto de Ciências Matemáticas e de Computação (ICMC), Ciências de Computação e Matemática Computacional - Universidade de São Paulo.

Maletzke, A. G., Lee, H. D., Zalewski, W., Matsubara, E.T., Coy, C. S. R., Fagundes, J. J., Goes, J. R. N., Chung, W. F., Mapeamento de informações médicas descritas em formulários para bases de dados estruturadas, VII Workshop de Informática Médica, 2007, pp. 1–10.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Rezende, S.O. (2003). *Sistemas Inteligentes – Fundamentos e Aplicações*. Editora Manole.

Sung, J., Bang, S. Y., Choi, S. (2006). A Bayesian network classifier and hierarchical Gabor features for handwritten numeral recognition, *Pattern Recognition Letters* (27), No. 1, pp. 66-75.

Zalewske, W., Lee, H. D., Voltolini, R. F., Maletzke, A. G., Fagundes, J. J., Góes, J., Coy, C. S. R., Wu, F. S. (2007) Reconhecimento de caracteres manuscritos para o mapeamento de formulários médicos para bases de dados estruturadas, X Congresso Brasileiro de Informática em Saúde, pp. 1-10.