Avaliação de Técnicas de Extração de Características para o Reconhecimento de Caracteres Manuscritos aplicado ao Mapeamento de Informações em Formulários Médicos

Willian Zalewski[1], Huei Diana Lee[1], André Gustavo Maletzke[1, 3], Newton Spolaôr[1], Adewole Caetano[4], Ana Carolina Lorena[4], João J. Fagundes[2], Cláudio S. R. Coy[2], Feng C. Wu[1,2]

[1]*Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná*
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Foz do Iguaçu, Paraná, Brasil

[2]*Faculdade de Ciências Médicas – Universidade Estadual de Campinas*
Serviço de Coloproctologia
Campinas, São Paulo, Brasil

[3]*Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo*
Laboratório de Inteligência Computacional – LABIC
São Carlos, São Paulo, Brasil

[4]*Centro de Matemática, Computação e Cognição – Universidade Federal do ABC*
Santo André, São Paulo, Brasil

Resumo
Pesquisas relacionadas a doenças envolvem, em geral, a análise de uma grande quantidade de dados. Nesse contexto, processos computacionais são cada vez mais utilizados para a extração e a análise de padrões que podem estar contidos nesses dados. No entanto, as informações são freqüentemente registradas em documentos impressos, impossibilitando a aplicação direta desses processos. Neste trabalho são apresentados a metodologia e o sistema computacional desenvolvidos para auxiliar no processo de mapeamento de formulários impressos para bases de dados. Também são descritas e avaliadas as técnicas de reconhecimento de caracteres empregadas e um estudo de caso sobre a doença de Crohn.

Palavras – Chave
Inteligência Artificial, Reconhecimento de Caracteres Manuscritos e Doença de Crohn.

1. Introdução

A análise de dados é uma tarefa que, cada vez mais, tem sido aplicada em diversas áreas com o intuito de auxiliar no processo de tomada de decisão. Todavia, o aumento do volume de dados armazenados tem tornado essa tarefa crescentemente complexa por meio de abordagens tradicionais e/ou manuais. Nesse sentido, cada vez mais métodos e processos computacionais têm sido propostos e aplicados na análise de grandes conjuntos de dados. Dentre esses processos, o de Descoberta de Conhecimento em Bases de Dados (DCBD) [1] tem sido utilizado como apoio na tarefa de análise de dados. Esse processo tem como

objetivo a extração de padrões contidos nos dados, de modo que esses padrões constituam uma fonte de informação interessante e relevante para especialistas de diversos domínios.

Para que o processo de DCBD possa ser aplicado é necessário que os dados estejam dispostos em um formato estruturado como a representação atributo-valor. No entanto, os dados encontram-se, geralmente, em formatos desestruturados ou semi-estruturados. Na área de medicina, os dados estão, freqüentemente, dispostos em laudos médicos e formulários impressos contendo informações como histórico do paciente e sintomatologia, fato este que dificulta a análise direta por meio de métodos computacionais. Diversos motivos estão relacionados à indisponibilidade desses dados em um formato adequado, dentre os quais a não existência de computadores em ambulatórios médicos, a consideração por muitos profissionais da área de medicina de que a utilização de documentos impressos torna o relacionamento com o paciente menos impessoal e a necessidade de se manter registros impressos. Portanto, para que processos, como o DCBD, possam ser aplicados é necessário representar esses dados em formato estruturado, como a representação atributo-valor [2, 3].

O objetivo deste trabalho é apresentar um sistema computacional que implementa uma metodologia para o mapeamento de dados a partir de formulários médicos impressos para bases de dados estruturadas e a avaliação de diferentes métodos para a Extração de Características (EC), por meio da construção e a avaliação de modelos para o Reconhecimento de Caracteres Manuscritos (RCM) contidos nesses formulários. Para avaliar a metodologia como um todo, é também apresentado um estudo de caso da aplicação dessa metodologia ao tema da doença de Crohn.

O restante deste trabalho está organizado do seguinte modo: na Seção 2 é apresentada a metodologia para o mapeamento de dados a partir de formulários médicos impressos para bases de dados estruturadas e na Seção 3 é apresentado o sistema no qual foi implementada essa metodologia; na Seção 4 são apresentados métodos para a EC para a representação das imagens dos caracteres manuscritos, bem como paradigmas comumente aplicados na área de reconhecimento de caracteres; a avaliação experimental realizada para determinar-se qual seria o melhor modelo para o reconhecimento de caracteres é descrita na Seção 5; na Seção 6 é apresentado o estudo de caso considerando o domínio da doença de Crohn e na Seção 7 são apresentadas as conclusões e trabalhos futuros.

2. Metodologia para o Mapeamento Automático de Formulários

Este trabalho está inserido no projeto de Análise Inteligente de Dados aplicado ao Mapeamento de Dados (AidMD), o qual está sendo desenvolvido pelo Laboratório de Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE) em conjunto com o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (UNICAMP), o Laboratório de Inteligência Computacional (LABIC) da Universidade de São Paulo (USP) em São Carlos, e o Centro de Matemática, Computação e Cognição (CMCC) da Universidade Federal do ABC (UFABC).

A metodologia proposta para o mapeamento de formulários médicos foi aplicada a alguns estudos de caso com sucesso [4, 5]. Essa metodologia foi organizada em três etapas: (1) Geração de Formulários e Construção da Base de Dados, (2) Construção de Padrões sobre Formulários e (3) Mapeamento de Formulários e Preenchimento da Base de Dados. Na primeira etapa são construídos formulários contendo campos de múltipla escolha e

campos numéricos, a partir dos atributos definidos anteriormente. Ainda outras características estão presentes no formulário como marcas de referência e identificação do formulário, necessárias nas próximas etapas. Na segunda etapa, o formulário gerado na etapa anterior é digitalizado e submetido a algoritmos de Inteligência Artificial (IA), os quais auxiliam na construção de padrões sobre esse formulário, de modo que o usuário não necessite preocupar-se com a declaração explícita da localização dos campos a serem mapeados. É importante ressaltar que essa etapa torna o processo de mapeamento de diferentes formulários, seja por ruídos causados no processo de digitalização ou pela distribuição dos campos, mais robusto e eficiente. Na terceira etapa cópias preenchidas do formulário são mapeadas para a Base de Dados (BD) por meio dos padrões gerados na etapa anterior.

3. Sistema de Mapeamento de Formulários – ForMappSys

O *Form Mapping System* (ForMappSys) consiste em um aplicativo computacional desenvolvido com o objetivo de integrar todas as etapas da metodologia de mapeamento de formulários. O sistema ForMappSys foi desenvolvido em linguagem JAVA[1] baseando-se no paradigma de programação orientado a objetos. As três etapas da metodologia para o mapeamento de formulários foram elaboradas e implementadas em quatro módulos dentro do sistema ForMappSys, os quais são:

Módulo de Geração de Formulários: por meio desse módulo são construídos os formulários a partir de um conjunto de perguntas e respostas, uma BD para a qual essas respostas serão mapeadas e um Arquivo de Interpretação (AI), o qual contém as regras de preenchimento da BD para cada tipo de formulário construído. O AI está representado na linguagem *Extensible Markup Language*[2] (XML) e é responsável por indicar o valor que deve ser armazenado na BD de acordo com as respostas que foram marcadas no formulário. Um exemplo de formulário é apresentado na Figura 1. Cada pergunta representa um atributo na BD e a resposta preenchida, o valor que é conferido a este atributo na BD. Após a definição das perguntas e suas respectivas respostas, os formulários são gerados de maneira automática em formato *Portable Document Format* (PDF). Os formulários gerados possuem uma Marca de Referência (MR), representada por uma linha horizontal, que se estende por quase a totalidade da largura no cabeçalho do formulário. Essa MR é utilizada para auxiliar na correção de imperfeições dos formulários e localização dos campos de preenchimento. Esse módulo disponibiliza uma interface gráfica (Figura 2 (a)) com o usuário que possibilita a edição de perguntas, respostas e informações que irão compor o formulário. Como mencionado, nessa etapa também é realizada a construção da BD, a qual foi implementada utilizando o Sistema de Gerenciamento de Banco de Dados MySQL[3].

Módulo de Construção de Padrões sobre Formulários: o objetivo desse módulo é mapear a localização dos campos a partir de um formulário modelo (não preenchido). O registro das informações sobre a localização dos campos do formulário é realizado em um arquivo denominado Base de Padrões. Por meio da interface gráfica (Figura 2 (b)) desenvolvida para esse módulo é possível visualizar a imagem do formulário modelo, as informações como a quantidade de campos de exemplos escolhidos, a quantidade de

[1] http://www.sun.com
[2] http://www.w3c.org/XML
[3] http://www.mysql.com

campos existentes no formulário, o número de campos que já foram reconhecidos no formulário e selecionar os campos que deverão ser mapeados.

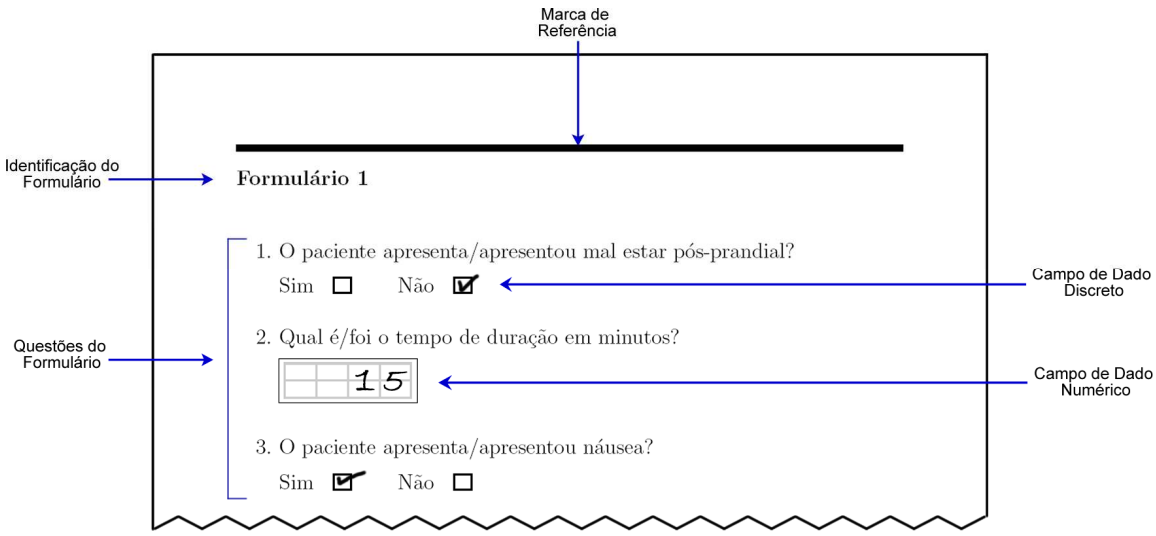

Figura 1 - Exemplo de Formulário.

Figura 2 - Interfaces Gráficas dos Módulos do Sistema ForMappSys.

Módulo de Mapeamento de Formulários: esse módulo tem como finalidade identificar qual resposta de cada pergunta foi preenchida no formulário e mapear essas informações para a BD, considerando as informações registradas na Base de Padrões. O Arquivo de Interpretação, gerado na primeira etapa da metodologia, é utilizado nesse momento para indicar, de acordo com o campo preenchido, qual será o valor a ser registrado na BD, exceto para os campos numéricos que são tratados por um módulo específico. Para esse módulo foi elaborada uma interface gráfica (Figura 2 (c)), a qual permite selecionar as imagens dos formulários que deverão ser mapeados para a BD, a Base de Padrões e o AI que serão utilizados. Também são disponibilizadas outras funcionalidades, como a seleção do modelo construído para a classificação que será considerado pelo Módulo de Reconhecimento de Caracteres Manuscritos.

Módulo de Reconhecimento de Caracteres Manuscritos: para a construção desse módulo, foram pesquisadas as principais técnicas adotadas na literatura para o reconhecimento de caracteres manuscritos. A partir desse estudo foram avaliadas diversas dessas técnicas de modo individual e combinadas. Com base nos resultados dessas avaliações foi construído um modelo de classificação utilizado por esse módulo. As técnicas utilizadas para o RCM e o processo de construção e avaliação dos modelos são apresentados nas Seções 4 e 5, respectivamente.

4. Reconhecimento de Caracteres Manuscritos

O problema do reconhecimento de caracteres escritos manualmente tem sido estudado há décadas e muitas técnicas foram propostas com o objetivo de resolver esse problema [6]. Nesse contexto, as estratégias geralmente adotadas pelos sistemas de reconhecimento de caracteres consistem nas seguintes etapas: (1) Pré-Processamento: minimização da existência de problemas que possam interferir de forma negativa no processo de reconhecimento; (2) Extração de Características: extração de informações a partir de dados brutos da imagem do caractere, formando um conjunto de características; (3) Classificação: consiste na aplicação de técnicas para determinar a qual classe pertence o caractere representado pelo conjunto de características obtidas na etapa anterior.

A EC constitui um fator de grande importância para que seja possível obter bons resultados em aplicações nessa área e apresenta como objetivo obter informações a partir da imagem de caracteres. Os métodos propostos para esse fim buscam minimizar a variabilidade dos padrões de uma mesma classe e ressaltar a diferença de padrões entre classes distintas e são, geralmente, classificados em duas diferentes categorias: características estruturais e características estatísticas [6, 7, 8, 9, 10].

Características Estruturais

Os métodos de EC Estruturais baseiam-se na análise estrutural da imagem. As informações obtidas correspondem a como os pixels de uma imagem estão arranjados na composição dos traços que constituem o caractere. Neste trabalho foram consideradas as seguintes características estruturais:

Pontos Finais: um pixel do caractere é considerado um ponto final, caso exista somente um pixel com valor igual a um em sua vizinhança. Desse modo, o método de Pontos Finais consiste em localizar e contabilizar a quantidade de pixels na imagem do caractere, os quais correspondem a pontos finais (Figura 3 (a));

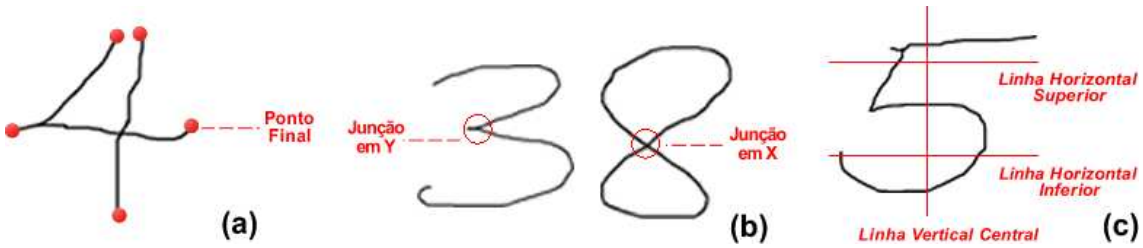

Figura 3 - Métodos de Extração de Características Estruturais.

Junções: as características extraídas referem-se ao número e à localização de junções em formato de *X* ou em formato de *Y* no caractere. Para obter essa informação, é aplicada sobre cada pixel que compõem a imagem do caractere, uma máscara 3×3, a qual possibilita identificar esses tipos de junções (Figura 3 (b));

Intersecções com Linhas Retas: consiste em identificar a quantidade e a posição de intersecções entre os traços que compõem o caractere e as linhas de referência sobrepostas na imagem. Essas linhas geralmente são organizadas em duas linhas horizontais e uma vertical para representar o caractere. As duas linhas horizontais são localizadas sobre a primeira e a segunda terça parte da altura da imagem do caractere, com o objetivo de prover um bom descritor as partes superiores e inferiores do caractere (Figura 3(c)).

Características Estatísticas

Os métodos de EC Estatísticas permitem obter informações de forma global a respeito da imagem do caractere, ou seja, o modo como estão distribuídos todos os pixels de um caractere. Os métodos considerados neste trabalho são apresentados a seguir.

Codificação Radial: baseia-se no fato de que o círculo é a única forma geométrica naturalmente invariante à rotação em uma imagem bidimensional. O algoritmo consiste em gerar uma quantidade *K* de círculos eqüidistantes ao redor do centro de massa da imagem (Figura 4 (a));

Perfis de Contorno: consiste em contabilizar a quantidade de pixels (distância) entre a borda da imagem do caractere e a borda do caractere (Figura 4 (b));

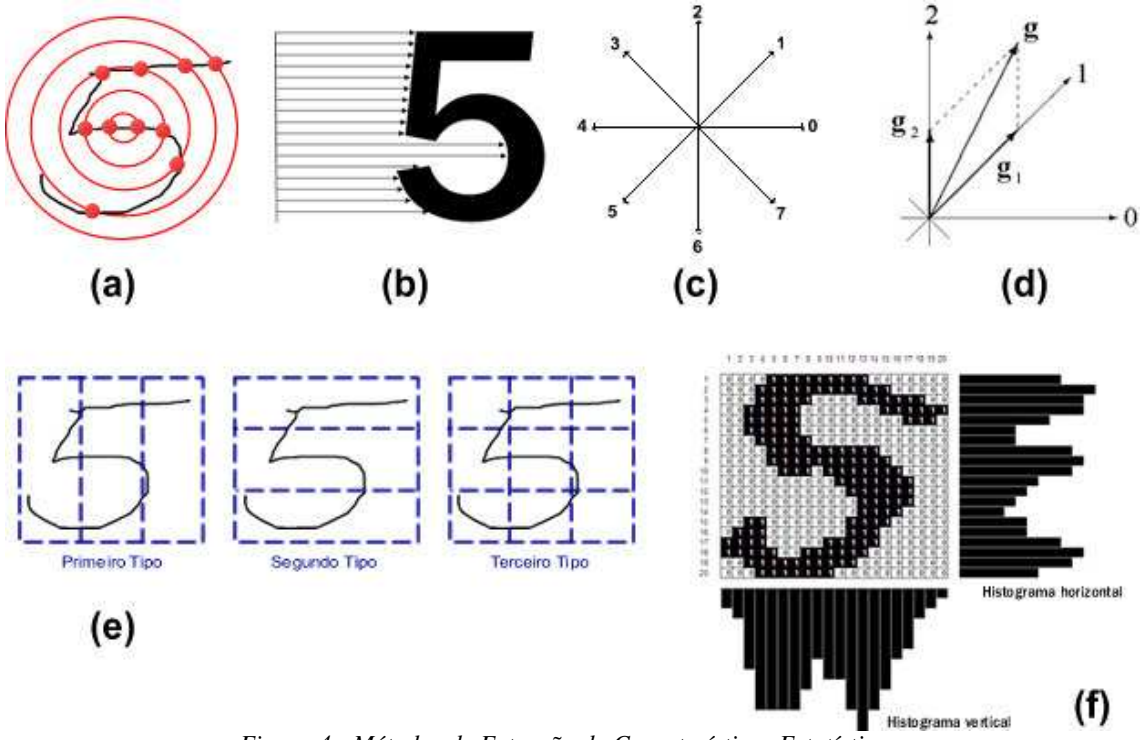

Figura 4 - Métodos de Extração de Características Estatísticas.

***Chaincode Direction*:** os pixels que compõem o contorno da imagem do caractere são agrupados em direções. A quantidade de direções comumente utilizada para esse método são quatro e oito direções. Um exemplo considerando oito planos de direções é apresentado na Figura 4 (c);

***Gradient Direction*:** utilizam-se duas máscaras do operador Sobel para computar o fator de inclinação em dois eixos para todos os pixels da imagem. As máscaras permitem que o fator de inclinação do pixel, em relação a seus vizinhos, seja extraído em duas componentes *x* e *y*, as quais são decompostas em planos direcionais de quatro ou oito direções. Caso o fator de inclinação calculado situar-se entre dois planos de direção, este será decomposto em duas componentes como representado na Figura 4 (d), onde *g* está entre os planos 1 e 2, e portanto foi decomposto para as duas componentes g_1 e g_2;

***Zoning*:** esse método consiste em dividir a imagem do caractere em um determinado número de zonas e para cada uma é contabilizada a quantidade de pixels

existentes. Um exemplo da aplicação desse método é apresentado na Figura 4 (e), no qual são utilizados três tipos de zoneamento. O primeiro baseia-se na divisão da imagem em três partições uniformes em relação à largura da imagem. O segundo tipo é semelhante ao primeiro, porém a divisão é realizada em função da altura da imagem. No terceiro tipo de aplicação do método de *Zoning* é utilizada uma união do primeiro e do segundo tipo, resultando em uma imagem particionada em nove regiões;

Histogramas de Projeção: o algoritmo consiste em contabilizar a quantidade de pixels para cada linha (histograma horizontal) ou coluna (histograma vertical) da imagem. Para obter o histograma de projeção horizontal, todas as linhas da imagem do caractere são analisadas e para cada pixel com conteúdo de valor igual a 1, a barra do histograma correspondente a essa linha é incrementada em uma unidade. O mesmo procedimento é realizado para gerar o histograma de projeção vertical, porém nessa estratégia as colunas é que são analisadas (Figura 4 (f)).

Classificadores

A Classificação constitui uma tarefa fundamental no processo de reconhecimento, de modo que todos os esforços aplicados nas etapas anteriores visam atingir melhores resultados nessa tarefa. Nesse contexto, na área de RCM, técnicas de Aprendizado de Máquina (AM) têm sido amplamente exploradas para a tarefa de Classificação. O AM é uma importante área em IA e tem como objetivo a construção de sistemas capazes de adquirir automaticamente conhecimento. Muitos métodos baseados nesses paradigmas de AM têm sido propostos para a construção de sistemas de RCM, dentre os quais destacam-se *k-Nearest Neighbor*, o qual classifica os exemplos de acordo com a similaridade em relação aos exemplos anteriormente vistos e *Multilayer Perceptron*, no qual um modelo composto por unidades simples, altamente interconectadas e organizadas em múltiplas camadas é utilizado para a classificação dos caracteres [7, 8, 11].

5. Construção e Avaliação Experimental do Modelo para RCM

Uma das tarefas essenciais para o desenvolvimento e a avaliação de métodos para RCM, é a comparação dos resultados por meio da utilização de alguma BD padrão de imagens de caracteres. Entre as diversas BD utilizadas como *benchmarks* para a realização desse tipo de avaliação, foi utilizada a base de imagens da MNIST [12]. A partir dessa base foram extraídos 1000 dígitos, sendo que para cada classe de dígito foram selecionadas 100 imagens aleatoriamente. Essas imagens, as quais encontram-se em escala de cinza, foram submetidas à aplicação da técnica de Binarização e posteriormente redimensionadas para um tamanho padrão de 45×35 pixels. Os classificadores utilizados para a construção dos modelos foram *k-Nearest Neighbor* e *Multilayer Perceptron* construídos com a ferramenta Weka [13] e considerando valores padrão. A avaliação dos modelos construídos foi realizada utilizando o método de Validação Cruzada com 10 partições. Para essa configuração, duas estratégias para a construção de modelos para classificação foram consideradas: (1) cada um dos métodos de EC, descritos na Seção 4, foi considerado individualmente para a construção de modelos e (2) os modelos foram construídos por meio da combinação de todos os métodos, a qual foi denominada de método Híbrido.

Resultados e Discussão

Os resultados são apresentados na Tabela 1, na qual é mostrada a precisão de cada método de EC em relação aos algoritmos de Classificação utilizados.

Atualmente, na área de RCM, é concordância da maioria dos pesquisadores que somente um simples método de EC não é, geralmente, suficiente para se alcançar bons resultados. Baseando-se nessa idéia, neste trabalho foi explorada a abordagem de combinação de múltiplos extratores de características (estruturais e estatísticas), selecionados de modo a representar diferentes propriedades dos caracteres. Essa abordagem provê uma completa descrição dos caracteres, pois a combinação dessas características possibilita minimizar as desvantagens de alguns métodos pelas propriedades de outros.

Tabela 1 - Resultados dos Experimentos.

Método de EC (1) Estrutural (2) Estatístico	***k-Nearest Neighbor* % Média (Desvio Padrão)**	***Multilayer Perceptron* % Média (Desvio Padrão)**
Hírido (1, 2)	97,67 (1,30)	99,29 (0,82)
Perfis de Contorno (2)	92,68 (2,37)	96,38 (1,85)
Histogramas de Projeção (2)	87,83 (3,39)	90,28 (2,68)
Zoning (2)	83,08 (3,58)	82,11 (3,72)
Chaincode Direction (2)	63,40 (4,92)	68,09 (4,12)
Gradient Direction (2)	61,95 (3,69)	72,87 (4,18)
Interseções com Linhas Retas (1)	61,55 (4,88)	69,13 (4,75)
Codificação Radial (2)	55,21 (4,70)	59,12 (4,24)
Pontos Finais (1)	47,31 (5,13)	55,87 (3,86)
Junções (1)	22,08 (1,30)	30,01 (3,89)

Por meio da aplicação do teste estatístico *t*, observou-se que o método Híbrido, foi o mais bem sucedido em ambos os algoritmos de Classificação (*k-Nearest Neighbor*: 97,67% e *Multilayer Perceptron*: 99,29%). Uma avaliação posterior mostrou que o modelo construído considerando o *Multilayer Perceptron* apresentou, estatisticamente, o melhor desempenho (*p-valor* < 0,001).

6. Estudo de Caso

Neste trabalho foi realizado um estudo de caso com o intuito de avaliar o modelo construído com a combinação proposta de características, utilizando um formulário construído com dados sobre a doença de Crohn, que tem despertado grande interesse entre os pesquisadores da área médica. Nessa doença inflamatória, células imunologicamente ativas agridem o aparelho digestivo, da boca até o ânus, em especial a parte inferior do intestino delgado (ílio) e do intestino grosso (cólon), provocando graves lesões como esfoliação, diarréia, dificuldade para absorver os nutrientes e enfraquecimento entre outros [14]. Atualmente, ainda não se conhece a causa exata da doença de Crohn; inúmeras pesquisas tentaram relacionar fatores genéticos, ambientais, alimentares e infecciosos como responsáveis pela doença. Porém, nenhum desses fatores, isoladamente, conseguiu explicar integralmente a etiologia e o padrão de desenvolvimento dessa doença.

Nesse contexto, processos computacionais como o de DCBD, poderiam ser úteis como ferramentas de apoio para pesquisas relacionadas a diversas enfermidades como a doença de Crohn. Isso permitiria que dados de acompanhamento de pacientes juntamente com os resultados de tratamentos pudessem ser utilizados para a construção de modelos e extração de padrões sobre essas enfermidades. No entanto, devido à grande quantidade de atributos que geralmente devem ser coletados para estudo sobre essas doenças, surgiu a necessidade de se elaborar um método eficiente e menos dispendioso para realizar a coleta desses dados. A metodologia apresentada neste trabalho visa auxiliar nesse aspecto.

Configuração Experimental

Com o apoio do sistema ForMappSys foi construído um formulário modelo, não preenchido, composto por cinco perguntas com respostas de múltipla escolha e quatro perguntas com respostas numéricas, o qual foi utilizado para a construção de padrões referentes à localização dos campos. Esse formulário foi impresso em escala de cinza utilizando uma impressora HP Color LaserJet 2605dn com 1200 pixels por polegada. Posteriormente, o formulário modelo foi digitalizado utilizando um *scanner* HP ScanJet 5550c com configuração de 75 pixels por polegada, 50% de brilho e 50% de contraste.

A partir do formulário modelo foram geradas 50 cópias, as quais foram preenchidas por dez colaboradores utilizando caneta esferográfica de cor preta. Para os campos de múltipla escolha o preenchimento foi realizado de modo *ad libitum*, apenas fixando-se como critério a realização do preenchimento próximo ao centro de cada marca e aplicando-se uma pressão normal de escrita. Em relação aos campos numéricos, foi recomendado o preenchimento dos caracteres dentro do espaço delimitado pela grade em tons de cinza, contida no campo numérico, e que os caracteres fossem escritos separadamente, de modo que não existisse conexão entre os mesmos. Após, os 50 formulários foram digitalizados seguindo a mesma configuração utilizada para a digitalização do formulário modelo. Os dados contidos nesses 50 formulários foram mapeados para a BD considerando para o reconhecimento dos caracteres manuscritos o modelo *Multilayer Perceptron*, apresentado e avaliado na Seção 5.

Resultados e Discussão

Considerando todos os formulários, em relação às 250 perguntas com respostas de múltipla escolha, obteve-se uma precisão de 99,60%, pois somente um campo não foi mapeado corretamente. Em relação ao reconhecimento dos 504 caracteres preenchidos nos campos numéricos, entre todos os formulários, foi constatada uma precisão de 96,23%. Os especialistas do domínio consideram os resultados promissores. A análise desses resultados mostrou que grande parte dos caracteres preenchidos nos formulários foram classificados corretamente. Um fator determinante no sucesso de aplicações que utilizam técnicas de RCM refere-se ao estabelecimento de restrições. Desse modo, é possível reduzir a grande variabilidade de parâmetros existentes nesse tipo de aplicação, permitindo uma maior confiabilidade em relação à veracidade dos dados.

7. Conclusão e Trabalhos Futuros

Os resultados obtidos neste trabalho foram considerados consistentes e incentivadores pelos especialistas do domínio, de modo que o custo de tempo e a subjetividade foram eliminados ou minimizados durante o mapeamento. Desse modo, trabalhos futuros incluem a seleção de atributos relevantes a partir dos métodos de Extração de Características e a utilização conjunta de técnicas de Classificação com o objetivo de aumentar a confiabilidade no reconhecimento dos caracteres numéricos preenchidos. Posteriormente, outro trabalho inclui a aplicação de métodos computacionais para a extração de padrões que possam estar contidos na Base de Dados, relacionada à doença de Crohn, construída por meio da metodologia de mapeamento de formulários.

Agradecimentos
Ao Programa de Desenvolvimento Tecnológico Avançado (PDTA) da Fundação Parque Tecnológico Itaipu (FPTI/BR) pelo auxílio no financiamento de bolsas.

Referências

[1] U. M. Fayyad, G. Platestsky-Shapiro e P. Smyth, "From data mining to knowledge discovery: an overview". *American Assoiation for Artificial Intelligence*, p. 1–30, 1996.

[2] F. D. Honorato, E. A. Cherman, H. D. Lee, M. C. Monard e F. C. Wu, Construcão de uma representação atributo-valor para extração de conhecimento a partir de informações semi-estruturadas de laudos médicos. *Conferencia Latinoamericana de Informática*, p. 1–12, Costa Rica, 2007.

[3] H. D. Lee, Seleção de atributos importantes para a extração de conhecimento de bases de dados, Tese de Doutorado, ICMC - USP, São Carlos, Brasil, 2005.

[4] A. G. Maletzke, H. D. Lee, F. C. Wu, E. T. Matsubara, C. S. R. Coy, J. S. Fagundes e J.R.N. Góes, Uma metodologia para auxiliar no processo de mapeamento de formulários médicos para bases de dados estruturadas. *X Congresso Brasileiro de Informática em Saúde*, p. 1-10, Florianópolis, Brasil, 2006.

[5] A. G. Maletzke, H. D. Lee, W. Zalewski, E. T. Matsubara, C. S. R. Coy, J. S. Fagundes, J.R.N. Góes e F. C. Wu, Mapeamento de informações médicas descritas em formulários para bases de dados estruturadas. *VII Workshop de Informática Médica*, p. 1-10, Porto de Galinhas, Brasil, 2007.

[6] O. D. Trier, A. K. Jain e T. Taxt, "Feature extraction methods for character recognition - a survey". *Pattern Recognition*, vol. 29, p. 641–662, 1996.

[7] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier e C. Olivier, "A structural/ statistical feature based vector for handwritten character recognition". *Pattern Recognition Letters*, vol. 19, p. 629–641, 1998.

[8] N. Arica e F. Yarman-Vural, An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems Man and Cybernetics Part C:Applications and Reviews*, vol. 31, p. 216–233, 2001.

[9] A. Koerich, Unconstrained handwritten character recognition using diferent classification strategies. *International Workshop on Artificial Neural Networks in Pattern Recognition*, p. 1–5, Firenze, Itália, 2003.

[10] C. L. Liu, K. Nakashima, H. Sako e H. Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques". *Pattern Recognition*, p. 265–279, 2004.

[11] S. Ouchtati, B. Ouchtati e A. Lachouri, "Segmentation and recognition of handwritten numeric chains". *International Journal of Informatics Technology*, vol. 4, p. 37–44, 2007.

[12] Y. Lecun, L. Bottou, Y. Bengio e P. Haffner, Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, p. 2278–2324, 1998.

[13] I. H. Witten e E. Frank, *Data Mining: practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, USA, 2 ed., 2005.

[14] R. S. Cotran, V. Kumar e T. Collins, *Robbins Patologia Estrutural e Funcional*, Guanabara Koogan, Rio de Janeiro, Brasil, 6 ed., 2000.

Datos de Contacto:

Willian Zalewski[1] – willzal@gmail.com
Huei Diana Lee[2] – hueidianalee@gmail.com
André Gustavo Maletzke[3] – andregm@icmc.usp.br
Newton Spolaôr[4] – newtonspolaor@gmail.com
Adewole Caetano[5] – adewole.caetano@ufabc.edu.br
Ana Carolina Lorena[6] – ana.lorena@ufabc.edu.br
Cláudio Saddy Rodrigues Coy[7] – ccoy@terra.com.br

João José Fagundes[8] – jjfagundes@mpcnet.com
Feng Chung Wu[9] – wufc@pti.org.br

[1,2,3,4,9]*Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná, Laboratório de Bioinformática – LABI, Parque Tecnológico Itaipu – PTI, Caixa Postal 39, 85856-970 – Foz do Iguaçu, Paraná, Brasil.*

[7,8,9]*Faculdade de Ciências Médicas – Universidade Estadual de Campinas, Serviço de Coloproctologia, Caixa Postal 6111, 13083 – 970 – Campinas, São Paulo, Brasil.*

[3]*Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, Laboratório de Inteligência Computacional, Caixa Postal 668,13560 – 970, São Carlos, São Paulo, Brasil.*

[5,6]*Centro de Matemática, Computação e Cognição – Universidade Federal do ABC, Caixa Postal 09,210-170, Santo André, São Paulo, Brasil.*