DETECÇÃO DE SPOTS EM IMAGENS ORIUNDAS DE GÉIS DE ELETROFORESE BIDIMENSIONAL UTILIZANDO CLUSTERIZAÇÃO FUZZY

MARLON DA SILVA DIAS[1]; GEANCARLO SANTANA MAYDANA[2]; MARILTON SANCHOTENE DE AGUIAR[3]

[1] *Universidde Federal de Pelotas – mdsdias@inf.ufpel.edu.br*
[2] *Universidde Federal de Pelotas – gsmaydana@inf.ufpel.edu.br*
[3] *Universidde Federal de Pelotas – marilton@inf.ufpel.edu.br*

1. INTRODUÇÃO

Proteínas têm um papel importante na vida e saúde, por conta disso é preciso compreendê-las, assim como estudar como elas operam no nível celular para assim entender como, por exemplo, regular os mecanismos de uma doenças (KLUG et al., 2011). A proteômica é definida como a caracterização em larga escala do conjunto de proteínas em uma célula ou tecido (HIRSCH et al., 2004). Seu principal objetivo consiste de documentar, identificar e caracterizar proteínas, além de elucidar as suas associações e funções (CIERO; BELLATO, 2002).

Ultimamente, uma das principais técnicas usadas na proteômica é a eletroforese bidimensional (KAVOOSI; ARDESTANI, 2012), a qual é baseada na separação e migração de moléculas, posicionadas em uma solução, sob influência da aplicação de um campo elétrico. Estas proteínas podem ser detectadas por uma variedade de reagentes de revelação, observando-se um perfil bidimensional de pontos. Ao final do processo, o gel resultante é escaneado e a imagem resultante pode ser processada.

As imagens de eletroforese bidimensional podem conter ruídos, assim como partículas de poeira e rachaduras no gel, e esses fatores podem interferir no resultado final da análise de reconhecimento dos *spots* (SAVELONAS; MYLONA; MAROULIS, 2012). A análise dessas imagens costumava a ser feito manualmente. Entretanto, o processo manual de avaliação da imagem é monótono, subjetivo, e propenso a erros, visto que depende fortemente da experiência individual do usuário no reconhecimento de spots (PARK et al., 2012).

Tendo em vista todos os pontos levantados anteriormente, esse trabalho visa propor o desenvolvimento de um sistema que utilize clusterização fuzzy para automatizar a realização dessa avaliação. Assim, tem-se por objetivo propor um modelo baseado em clusterização fuzzy para realizar o reconhecimento de *spots* em imagens de géis de eletroforese bidimensional.

2. METODOLOGIA

Para a realização da detecção de *spots* escolheu-se a clusterização fuzzy, através do algoritmo conhecido como Fuzzy C-Means. Os conjuntos fuzzy são o modelo mais tradicional para o tratamento de informações vagas e inexatas. Introduzido por ZADEH (1965) tem como objetivo permitir um elemento pertencer, com um certo grau de intensidade, a uma dada classe. A representação com conjuntos fuzzy utiliza conjuntos para representar a informação que não é precisa e emprega lógica fuzzy para a tomada de decisão, provendo um mecanismo para representar e manipular algum tipo de incerteza e ambiguidade.

O objetivo da análise da clusterização é dividir os dados de entrada em grupos, chamados de *clusters*, de tal forma que, com base um uma métrica, os membros de um mesmo grupo são mais parecidos entre si do que com os

membros de outros grupos (MURPHY, 2012). No processo de clusterização não-fuzzy, cada amostra é atribuída a somente um cluster e todos os clusters são conjuntos disjuntos.

Na prática, entretanto, existem muitos casos em que os clusters não são completamente disjuntos e os dados podem ser classificados como pertencendo mais a um cluster que a outro. Assim, a separação dos clusters traz uma noção fuzzy, e as representações de estruturas de dados reais podem então serem tratadas com mais precisão por métodos de agrupamento fuzzy.

O algoritmo Fuzzy C-means (FCM) é o mais conhecido e a técnica de clusterização fuzzy mais utilizada (CHI; YAN; PHAM, 1996). Este algoritmo é desenvolvido baseado na minimização iterativa da seguinte função critério:

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}|x_k - v_i|^2$$

onde

- $x_1, \cdots, x_n$ representa os n dados a serem clusterizados;
- $V = \{v_1, \cdots, v_n\}$ representa os centros dos clusters;
- $U = [u_{ik}]$ é uma matriz $c \times n$, onde u_{ik} é o i-ésimo grau de pertinência do k-ésimo dado de entrada x_k, e o grau de pertinência satisfaz as seguintes condições

$$0 \leq u_{ik} \leq 1, \quad i = 1,2,\cdots,c; k = 1,2,\cdots,n$$

$$\sum_{k=1}^{c} u_{ik} = 1 \quad k = 1,2,\cdots,n$$

$$0 < \sum_{k=1}^{n} u_{ik} < n \quad i = 1,2,\cdots,c$$

- $m \in [1,\infty)$ é um fator de peso expoente.

A função objetivo é a soma das distâncias Euclidianas entre as amostras e seus centro de cluster correspondente, elevada ao quadrado, com as distâncias sendo ponderadas pela pertinência fuzzy. No cálculo de um centro de cluster, todas as amostras de entrada são consideradas, sendo elas ponderadas pelos seus valores de pertinência. Para cada amostra, seu valor de pertinência em cada classe depende de sua distância ao centro de cluster correspondente. O fator m reduz a influência de pequenos valores de pertinência. Quanto maior o valor de m, menor é a influência de amostras com menor valor de pertinência.

O FCM consiste dos seguintes passos (CHI; YAN; PHAM, 1996):

1. Inicialize $U^{(0)}$ aleatoriamente ou baseado em uma aproximação; inicialize $V^{(0)}$ e calcule $U^{(0)}$. Defina o contador de iteração $\alpha = 1$. Defina o número de clusters c e defina o peso m.
2. Compute o centro dos clusters. Dado $U^{(\alpha)}$, calcular $V^{(\alpha)}$ segundo

$$v_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^{m}} \sum_{k=1}^{n} u_{ik}^{m}\, x_{ik} \quad i = 1,2,\cdots,c$$

3. Atualizar os graus de pertinência. Dado $V^{(\alpha)}$, calcular $U^{(\alpha)}$ segundo

$$u_{ik} = \frac{\left[\frac{1}{|x_k - v_i|^2}\right]^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left[\frac{1}{|x_k - v_j|^2}\right]^{\frac{1}{m-1}}} \quad i = 1,2,\cdots,c;\ k = 1,2,\cdots,n$$

4. Para a iteração se

$$\max\left|u_{ik}^{(\alpha)} - u_{ik}^{(\alpha-1)}\right| \leq \varepsilon$$

caso contrário, define $\alpha = \alpha + 1$ e volta para o Passo 2, onde ε é um valor definido previamente como sendo a menor mudança aceitável em U.

3. RESULTADOS E DISCUSSÃO

O modelo proposto visa fazer a clusterização das imagens com o intuito de encontrar *spots*, utilizando o algoritmo de clusterização Fuzzy C-means. A análise é feita com base na imagem provida ao final do exame, onde as moléculas já sofreram os efeitos dos campos elétricos e pararam de se mover. Partindo da imagem, os únicos dados presente são os tons de cinza que compõem a imagem. Baseando-se nos princípios da clusterização, acredita-se que os *spots* apresentam tons de cinza simulares e, assim, pertençam ao mesmo cluster.

Em HOOGLAND et al. (2004) é apresentado uma base de imagens oriundas de eletroforese bidimensional. As imagens já foram testadas e apresentam *spots* confirmados, os quais foram usados para medir a precisão do modelo. Logo, há uma lista informando os pontos na imagem que representam um *spot*. Os pontos apresentados nessa lista são *spots* confirmados, entretanto, ainda há outros pontos na imagem que não foram confirmados *spots*.

Diversos testes foram realizados com diferentes configurações do modelo. O resultado mais promissor foi utilizando três clusters. Com base nessa configuração, observa-se os clusters representando o fundo da imagem, redondezas ou rastros de um *spot* e, por fim, os *spots*. O modelo informa o grau de pertinência de cada ponto a um dos clusters. A função máximo é aplicada como método de defuzzificação, ou seja, o cluster é designado com base no maior grau de pertinência. Assim, tem-se a informação do cluster de cada ponto. Para poder contar um *spot* é necessário saber a área que representa uma *spot*, o conjunto de pontos próximo pertencentes ao cluster que representa os *spots*.

Um dos testes realizados foi em uma imagem chamada de ECOLI. Nesse teste, a lista de *spots* confirmados contém 206 *spots*. O modelo encontrou 279 *spots*, onde 151 estão na lista de confirmados, e 55 estão faltando, resultando em uma precisão de 73,3%. Os pontos não confirmados não diz necessariamente significam que não são *spots*. Na verdade, são áreas que poderiam ser consideradas *spots*, indicando que elas precisam de uma análise mais detalhada. De acordo com o protocolo do exame, pode ser necessário recortar uma porção do gel para uma análise mais detalhada e, assim, esta região em que o modelo indica ter spots não comprovados poderia ser uma sugestão de recorte.

4. CONCLUSÕES

Imagens de eletroforese bidimensional são difíceis de analisar. Há o problema de possíveis ruídos na imagem, o que pode dificultar na análise. Porém, além disso, os *spots* podem apresentar características bem diferentes. Por exemplo, dos *spots* confirmados na lista, mas não encontrados pelo modelo, apresentam tons de cinza muito distintos. A maioria dos confirmados possui um tom mais forte, entretanto, alguns dos confirmados possuem tons de cinza bem baixos. Isso é um grande desafio a ser superado pela clusterização.

Como trabalhos futuros, é previsto a utilização de algumas técnicas de pré-processamento. É possível utilizar algumas técnicas como remoção do fundo,

normalização, equalização e segmentação da imagem. As imagens utilizadas são relativamente grandes, contendo muita informação. A primeira técnica tem o intuito de reduzir os dados a serem processados, o que pode acelerar o processo. A segunda e terceira, visam mudar os dados apresentados, podendo influenciar diretamente na maneira com que a função fuzzy os clusteriza. E, por fim, a última técnica, a segmentação da imagem, tem como objetivo dividir a imagem, resultando na clusterização de imagens menores. A imagem original apresenta regiões com intensidades diferentes, com isso pretende-se achar aqueles *spots* com tons de cinza mais discrepantes.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CHI, Z; YAN, H.; PHAM T. **Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition**. USA: World Scientific, 1996.

CIERO, L. D.; BELLATO, C. d. M. Proteoma. **Biotecnologia Ciência & Desenvolvimento**, Brasília,v.29, p.158-164, 2002.

HIRSCH, J.; HANSEN, K. C.; BURLINGAME, A. L.; MATTHAY, M. A. Proteomics: current techniques and potential applications to lung disease. **American Journal of Physiology - Lung Cellular and Molecular Physiology**, USA, v.287, n.1, p.L1–L23, 2004.

HOOGLAND, C.; MOSTAGUIR, K.; SANCHEZ, J.; HOCHSTRASSER, D. F.; APPEL, R. D. SWISS-2DPAGE, ten years later. **Proteomics**, Basel, Switzerland, v.4, n.8, p.2352–2356, 2004.

KLUG, W. S.; CUMMINGS, M. R.; SPENCER, C. A.; PALLADINO, M. A. **Concepts of Genetics**. USA: Benjamin Cummings, 2011.

KAVOOSI, G.; ARDESTANI, S. K. Gel Electrophoresis of Protein – From Basic Science to Practical Approach. In: MAGDELDIN, S. (Ed.). **Gel Electrophoresis – Principles and Basics**. USA: InTech, 2012. Cap.6, p.69–88.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. Cambridge: The MIT Press, 2012.

PARK, S. C.; NA, I. S.; HAN, T. H.; KIM, S. H.; LEE, G. S. Lane detection and tracking in PCR gel electrophoresis images. **Computers and Electronics in Agriculture**, USA, v.83, n.0, p.85 – 91, 2012.

SAVELONAS, M. A.; MYLONA, E. A.; MAROULIS, D. Unsupervised 2D Gel Electrophoresis Image Segmentation Based on Active Contours. **Pattern Recognition**, USA, v.45, n.2, p.720–731, 2012.

ZADEH, L. A. Fuzzy sets. **Information and Control**, USA, v. 8, n. 3, p. 338–353, 1965.