XIV CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL CAMPINA GRANDE

FUZZY KOHONEN CLUSTERING NETWORK PARA DADOS INTERVALARES COM DISTÂNCIA CITY-BLOCK PONDERADA

Danilo de Menezes Freitas [1] **Carlos Wilson Dantas de Almeida** [2]

RESUMO

Este trabalho tem como objetivo realizar avanços nas técnicas de agrupamento entre dados intervalares. Partindo deste preceito foi feita a utilização do MatLab para desenvolver progressos no algoritmo Fuzzy Kohonen Clustering Network utilizando a distância city-block, a fim de obter resultados relevantes a partir do uso desse tipo de distância. Para finalizar foram realizadas comparações entre o algoritmo proposto com outros modelos para testar a real eficiência do algoritmo trabalhado.

[1] Graduando em Ciência da Computação, DSC, UFCG, Campina Grande, PB, e-mail: danilomf14@gmail.com

[2] Doutor em Ciência da Computação, Professor, Departamento de Sistemas e Computação (DSC), UFCG, Campina Grande, PB, email: carlos.wilson@computacao.ufcg.edu.br

DESENVOLVIMENTO DE UM SISTEMA PARA REDUÇÃO DE RUÍDO PROVOCADO POR ARTEFATO METÁLICO EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA

ABSTRACT

This study aims to develop a system to make progress in the clustering techniques of interval data, with this in mind it was used MatLab to develop progress in the Fuzzy Kohonen Clustering Network algorithm using city-block distance in order to obtain relevant results from the use of this type of distance. To conclude, comparisons were made between the algorithm proposed with other models to test the real efficiency of the algorithm.

INTRODUÇÃO

Quando se pensa sobre o conceito de "informação", o que provavelmente vem à mente são longas sequências de símbolos ou caracteres. Sob esse aspecto, os sistemas computacionais são ótimas ferramentas para armazenamento, organização e recuperação.

Essa quantidade exuberante de dados que não para de aumentar vem surgindo desde a criação da internet. Cada vez mais se faz uso da internet, seja para lazer, como assistir a um filme, consumo, como comprar uma geladeira, serviços, como agendar uma consulta no dentista, e diversas outras coisas que só fazem com que a humanidade seja ainda mais dependente da internet e do armazenamento desses dados.

A mineração de dados é uma das formas de organização e seleção de grandes conjuntos de dados, o processo ocorre a partir de uma análise matemática para ser feito um reconhecimento de padrões ou tendências a fim de ordenar cada informação em seu grupo específico. Para isso existem diversos tipos de algoritmos, criados a partir de 6 diferentes etapas: definição do problema, preparação dos dados, exploração dos dados, criação de um modelo, validação do modelo e a implantação e atualização do mesmo.

Bezdek (BEZDEK; TSAO; PAL, 1992; TSAO; BEZDEK; PAL, 1994) desenvolveram o algoritmo *Fuzzy Kohonen Clustering Network* que é um algoritmo de agrupamento não-supervisionado que combina as ideias de valores de pertinência para as taxas de aprendizagem e o paralelismo do algoritmo *Fuzzy C-Means* (FCM) (BEZDEK, 1981) com as regras de atualização auto-organizáveis do algoritmo *Kohonen Clustering Network* (KCN) (KOHONEN, 2001). O objetivo do trabalho é avaliar e comparar o desempenho do algoritmo usando a distância euclidiana (versão original do algoritmo FKCN) e com a distância de cityblock. A ideia da distância de *City-Block* em vem de um trabalho com sucesso nesse sentido usano outro algoritmo (nuvem dinâmica) (SOUZA, 2003).

A distância de city-block é um tipo de distância assim como a euclidiana, que são abordados para calcular distâncias entre dois pontos específicos, porém os métodos utilizados para esses cálculos são diferentes. A distância euclidiana é calculada a partir de uma reta diagonal entre os dois pontos designada a partir do Teorema de Pitágoras, como se pode ver na Equação 1:

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \tag{1}$$

Já a distância city-block, fazendo uma analogia ao seu nome, remete a um “bloco” de uma vizinhança, onde sempre é considerado a distância em linhas retas,

como se fossem as ruas da cidade, andando apenas para cima, para baixo, esquerda e direita, não podendo ser considerada as diagonais como na distância euclidiana. Seu cálculo é feito como na Equação 2.

$$D(p,q) = |p_1 - q_1| + |p_2 - q_2| \tag{2}$$

MATERIAIS E MÉTODOS

Antes de começar a implementar o algoritmo Fuzzy Kohonen Clustering Network foi feito um estudo de diversos outros modelos posteriores a esse para um melhor entendimento do que é o algoritmo. De início, foi estudado o algoritmo de agrupamento k-means que tem como objetivo particionar n observações dentre k grupos (*clusters*) diferentes na qual cada observação pertence ao grupo em que sua distancia é mais perto da média daquele grupo, portanto a cada iteração do algoritmo os agrupamentos se tornam cada vez mais precisos até chegar em um ponto de convergência na qual as mudanças são quase imperceptíveis.

A partir de todo esse estudo foi feito sua implementação no MATLAB fazendo uso de dados sintéticos como teste. São predefinidos o número de clusters a serem agrupados, os pontos iniciais de cada cluster, o número de iterações e os dados a serem utilizados. A partir disso o algoritmo calcula a distancia de cada ponto no gráfico para o ponto de incial de cada n clusters existentes, o dado comparado é agrupado no cluster de menor distância, isso é feito com todos os dados. Após isso é calculada a média aritmética dos dados de cada cluster e assim o método se repete com mais uma iteração, até que se finalize o processo.

O algoritmo k-means é um dos modelos mais simples de agrupamentos de dados disponíveis, portanto sua eficiência comparada ao Fuzzy Kohonen Clustering Network é inferior. Esse que já foi testado com muito sucesso para distância euclidiana padrão, seria agora testado e analisado nesse projeto com a distância city-block a fim de amplificar o uso do modelo em questão, trazendo melhores resultados e análises.

No algoritmo FKCN (Fuzzy Kohonen Clustering Network), se trabalhou com um conceito diferente do k-means, A rede FKCN (TSAO; BEZDEK; PAL, 1994) é uma rede não-supervisionado que combina as ideias de valores de pertinência para as taxas de aprendizagem e o paralelismo do algoritmo *Fuzzy C-Means* (FCM) (BEZDEK, 1981) com as regras de atualização auto-organizáveis do algoritmo *Kohonen Clustering Network* (KCN) (KOHONEN, 2001).

Na rede KCN o treinamento é sequencial. Isso significa que a rede é atualizada após a apresentação de cada amostra. O conjunto de amostras é apresentado repetidas vezes à rede até que esta atinja a estabilidade.

O KCN sofre de alguns problemas. Para solucionar estes problemas, a rede FKCN foi criada, baseada na integração dos métodos FCM (BEZDEK, 1981) e do KCN (KOHONEN, 1989; KOHONEN, 2001). Neste novo método, a taxa de aprendizado é controlada automaticamente e com treinamento em lote (*batch*).

O término do algoritmo KCN ocorre no limite do número de iterações, a menos que a taxa de aprendizagem seja forçada para zero, o que é uma estratégia de terminação artificial. Já a terminação do algoritmo combinado FKCN ocorre antes que se atinja o número máximo de iterações. Bezdek completou a integração da FCM e KCN, definindo a taxa de aprendizagem de Kohonen como:

$$\begin{aligned} m_t &= m_0 - t * \Delta m \\ &= m_0 - t * \frac{m_0 - 1}{t_{max}} \end{aligned} \tag{3}$$

onde m_0 é o valor inicial do expoente de peso maior que um, t_{max} é o número máximo de iterações.

Bezdek (TSAO; BEZDEK; PAL, 1994) também prova que o FKCN é equivalente ao *fuzzy K-means* quando m é fixo, e ao *hard K-means* quando $m = 1$. Sua taxa de aprendizagem θ é definida como:

$$\theta_{ik,t} = (u_{ik,t})^{m_t} \tag{4}$$

onde $u_{ik,t}$ é o valor de pertinência *fuzzy* do padrão de entrada $\mathbf{x}_k$ no *i*-ésimo *cluster*. Quando m_t se aproxima de um, $u_{ik,t}$ pode ser zero ou um, como no modelo FCM. Os principais passos do FKCN estão descritos no Algoritmo 1.

```
%-------------------------------------------------------------
% Função: FKCN
%
% Entradas:
% v       - vetor de protótipos
% c       - número de agrupamentos
% t_max   - número máximo de iterações
% m_0     - valor de fuzzificação (sendo m_0 > 1)
% epsilon - parada
%
% Saídas:
% v_t     - vetor de pesos final
%-------------------------------------------------------------
function [v_t] = FKCN(v, c, t_max, m_0, ε)

  for t = 1 to t_max do

    m_t ← m_0 − t · [(m_0 − 1)/t_max]

    for k = 1 to N do
      for i = 1 to c do
        u_ik,t ← [ Σ_{h=1}^{c} ( Σ_{j=1}^{p} (x_k^j − v_{i,t−1}^j)^2 / Σ_{j=1}^{p} (x_k^j − v_{h,t−1}^j)^2 )^{1/(m_t − 1)} ]^{−1}

        θ_ik,t ← (u_ik,t)^{m_t}
      end
    end

    % Atualiza todos os protótipos
    for i = 1 to c do
      v_i,t ← v_i,t−1 + [ Σ_{k=1}^{n} θ_ik,t (x_k − v_i,t−1) ] / ( Σ_{k=1}^{n} θ_ik,t )
    end

    E_t ← ‖v_t − v_t−1‖^2

    if E_t ≤ ε do
      return v_t
    end

  end

end function
```

Algorithm 1: *Fuzzy Kohonen Clustering Network*.

RESULTADOS E DISCUSSÃO

Avaliamos o desempenho do algoritmo usando o índice corrigido de Rand (CR) (HUBERT; ARABIE, 1985). O índice CR calcula um grau de similaridade entre uma partição A e uma partição B fornecida pelo algoritmo FKCN.
O índice CR tem toma seus valores no intervalo de $[-1:1]$, em que o valor 1 indica o melhor valor obtido possível. Ou seja, quanto mais próximos de 1, melhor a classificação. Quando mais próximos de 0 ou valore negativos, menor a similaridade entre os grupos (JAIN; DUBES, 1988).

3.1 CONJUNTOS DE DADOS

Para fazer uma avaliação do algoritmo FKCN, vamos escolher três conjuntos de dados. São eles: **Peixes**, **Carro** e **Iris**. Os conjuntos estão disponíveis no repositório SODAS (*Symbolic Official Data Analysis System*) (DIDAY; NOIRHOMME-FRAITURE, 2008). Os conjuntos possuem atributos que identifica a classe de cada elemento da base de dados.
O experimento do algoritmo FKCN foi feita por base de um experimento Monte Carlo consideranto 100repetições para cada conjunto de dados reais. O objetivo de usar Monte Carlo é uma melhor avalisação do desempenho do método. Para os experimentos, o número de iterações foi fixado em 1000 e o valor inicial do expoente de peso $m_0 = 2$.

3.1.1 Conjunto de Dados Peixes

Estudos realizados na Guyana francesa têm indicado níveis anormais de contaminação de mercúrio em algumas regiões. Esta contaminação de mercúrio é devida ao alto índice de consumo de peixe de água doce contaminado (BOBOU; RIBEYRE, 1998). Com o objetivo de obter um melhor conhecimento deste fenômeno, um conjunto de dados foi coletado por pesquisadores do laboratório LEESA.

3.1.2 Conjunto de Dados Carros

O conjunto de dados simbólicos CARROS pode ser encontrado em Diday e Noirhome-Fraiture (DIDAY; NOIRHOMME-FRAITURE, 2008). Este conjunto possui 33 modelos de carros descritos por 8 variáveis (Preço, Cilindradas, Velocidade Máxima, Aceleração, Distância entre os Eixos, Comprimento, Largura, Altura). Cada indivíduo do conjunto será classificado como Utilitário, Berlina, Luxuoso e Esportivo.

3.1.3 Conjunto de Dados Iris

Outro conjunto clássico de classificação é o conjunto Iris (VESANTO et al., 1999). É um conjunto de flores do gênero Iris que são divididas em 3 grupos (rótulos): setosa, virginica e versicolor. O objetivo é determinar a qual grupo uma determinada flor pertence baseado nas medidas de sépalas e pétalas.

3.2 RESULTADOS

Nesta subseção são apresentados os resultados dos conjuntos de dados comparando o algoritmo FKCN com distância Euclidiano e Cityblock. Os resultados estão apresentados na Tabela 1.

3.2.1 Conjunto de Dados Peixes

Os índices de CR obtidos nos resultados do conjunto de dados peixes mostrado na Tabela 1 são, respectivamente, -0.1047 e -0.1047 para os algoritmos FKCN com distância euclidiano e com distância de cityblock. O número de iterações necessárias para convergir foram 958 e 961 para os algoritmos FKCN com distância euclidiana e com distância de cityblock. Em conclusão, para este conjunto de dados, não tivemos diferença no índice de CR, apenas no número máximo de iterações onde a versão com distância euclidiana convergiu em média 3 iterações a menos do que sua versão com distância cityblock.

3.2.2 Conjunto de Dados Carros

Os índices de CR obtidos nos resultados do conjunto de dados carros mostrado na Tabela X são, respectivamente, 0.2928 e 0.2648 para os algoritmos FKCN com distância euclidiano e com distância de cityblock. O número de iterações necessárias para convergir foram 967 e 989 para os algoritmos FKCN com distância euclidiana e com distância de cityblock. Em conclusão, para este conjunto de dados, a versão com distância euclidiana obteve um melhor resultado, tanto para o índice de CR quanto para o número de iterações do que sua versão com distância cityblock.

3.2.3 Conjunto de Dados Iris

Os índices de CR obtidos nos resultados do conjunto de dados Iris mostrado na Tabela X são, respectivamente, 0.7294 e 0.8399 para os algoritmos FKCN com distância euclidiano e com distância de cityblock. O número de iterações necessárias para convergir foram 17 e 13 para os algoritmos FKCN com distância euclidiana e com distância de cityblock. Em conclusão, para este conjunto de dados, a versão com

distância cityblock foi superior tanto para o índice de CR quanto para o número de iterações do que sua versão com distância cityblock.

Tabela 1 – Resultados Obtidos.

	CR (euclidiana)	CR (cityblock)	Número de iterações (euclidiana)	Número de iterações (cityblock)
Peixes	-0.1047	-0.1047	958	961
Carros	0.2928	0.2648	967	989
Iris	0.7294	0.8399	17	13

CONCLUSÃO

A partir dos estudos feitos nos algoritmos apresentados neste trabalho, pode-se notar a diferença entre os métodos hard e fuzzy de agrupamento de dados, consistindo principalmente na eficiência dos dois modelos.

No k-means, houve um trabalho mais direto com os dados sintéticos, e no FKCN houve a abordagem fuzzy, na qual obteve um desempenho superior usando a base IRIS e um desempenho inferior usando a base Peixes e Carros. Mais testes seriam necessários para determinar o desempenho do FKCN com distância City-block comparada com a distância Euclidiana.

Alguns dos objetivos planejados anteriormente no trabalho não puderam ser concluídos de forma integral, a implementação do algoritmo Fuzzy Kohonen Clustering Network usando a distância city-block para dados intervalares não pôde ser finalizada. Portanto as comparações que seriam feitas com o mesmo algoritmo porém com a distância euclidiana não pôde ser feita. Sendo assim, não pôde ser atestado a diferença e a relevância do uso do tipo de distância proposta.

Tendo como motivos para a não conclusão de alguns objetivos, pode-se citar as dificuldades que tive durante o trabalho, por alguma dificuldade na linguagem trabalhada, dificuldades pessoais, na falta de experiência e na minha ausência em responder as tentativas de contato do orientador na segunda metade do cronograma.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil no programa PIBIC/UFCG.

REFERÊNCIAS

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BEZDEK, J. C.; TSAO, E. C.-K.; PAL, N. R. Fuzzy kohonen clustering networks. In: **Proc. of the First IEEE Conference on Fuzzy Systems, 1992**. San Diego, USA: [s.n.], 1992.

BOBOU, A.; RIBEYRE, F. Mercury in the food web: Accumulation and transfer mechanisms. **Metal Ions in Biological Systems**, n. 34, p. 289–319, 1998.

DIDAY, E.; NOIRHOMME-FRAITURE, M. (Ed.). **Symbolic Data Analysis and the SODAS Software**. [S.l.]: Wiley-Interscience, 2008. ISBN 978-0470018835.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v. 2, n. 1, p. 193–218, 1985.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

KOHONEN, T. **Self-organization and associative memory: 3rd edition**. New York, NY, USA: Springer-Verlag New York, Inc., 1989. ISBN 0-387-51387-6.

KOHONEN, T. **Self-Organizing Maps**. 3rd edition. ed. [S.l.]: Springer-Verlag, 2001.

SOUZA, R. M. C. R. de. **Métodos de Cluster para Intervalos Usando Algoritmos do Tipo Nuvens Dinâmicas**. Tese (Doutorado) — Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), 2003.

TSAO, E. C.-K.; BEZDEK, J. C.; PAL, N. R. Fuzzy kohonen clustering networks. **Pattern Recognition**, v. 27, n. 5, p. 757–764, 1994.

VESANTO, J. et al. Self-organizing map in matlab: the SOM toolbox. In: **Proceedings of the Matlab DSP Conference**. Espoo, Finland: [s.n.], 1999. p. 35–40.