Uma Extensão Intervalar do Algoritmo Fuzzy C-Means

Rogério R. de Vargas*, **Benjamín R. C. Bedregal**
Programa de Pós-Graduação em Sistemas e Computação, DIMAp, UFRN
Lagoa Nova 59078-970 Natal, RN - Brazil
E-mail: rogerio@ppgsc.ufrn.br, bedregal@dimap.ufrn.br

Resumo: *Clusterização é o processo de organizar uma coleção de padrões em grupos baseados em suas similaridades. Técnicas de agrupamento fuzzy objetivam encontrar grupos aos quais todo objeto da base de dados pertence em algum grau de pertinência. Este trabalho apresenta uma extensão intervalar do algoritmo fuzzy c-means. Esta extensão possibilita que a entrada de dados e o grau de pertinência sejam intervalos. Possibilitando assim, representar os dados sem nenhuma conversão dos dados intervalares para pontuais.*

Palavras-chave: *Clusterização Fuzzy Intervalar, Processamento de Dados, Fuzzy C-Means, Dados Intervalares*

1 Introdução

Diariamente as pessoas armazenam ou representam grandes quantidades de dados com o propósito de análisa-los e manipulá-los posteriormente. Para aprender algo novo ou entender um fenômeno, é normal as pessoas sempre tentarem procurar as características que descrevam este objeto ou fenômeno e em seguida compará-los com outros objetos ou fenômenos conhecidos, baseado na similaridade ou dissimilaridade. Deste fato decorre o papel de sistemas de classificação ou de clusterização para a aquisição de conhecimento.

Análise de cluster é uma técnica aplicada a diversas áreas como a mineração de dados, reconhecimento de padrões e processamento de imagens. Algoritmos de clusterização têm por objetivo particionar um conjunto de dados em clusters de tal forma que tenham um alto grau de similaridade, enquanto indivíduos pertencentes a diferentes clusters tenham alto grau de dissimilaridade.

Nos métodos de Clusterização *hard* cada ponto no conjunto de dados pertence a exatamente um cluster. Clusterização Fuzzy de uma forma bem geral, tem a sua partição baseada na ideia de funções de pertinência expressa por um grau de pertinência referente a um cluster, isto é, os algoritmos fuzzy associam um dado a todos os clusters através da variação do grau de pertinência do dado em cada cluster.

O uso da representação da entrada dos dados na forma intervalar, garante a segurança na sua qualidade através do grau de incerteza, o diâmetro do intervalo, que pode ser visto como um indicativo da influência de erros nos dados a serem processados.

Em [1] foi proposto uma extensão intervalar do algoritmo fuzzy c-means, onde cada dado de entrada é um intervalo. Para calcular a distância de cada ponto a um determinado cluster (de intervalo) é usado a distância Euclidiana. Na finalidade de validar o método proposto, foi realizado vários testes em conjuntos de dados intervalares, um consistindo na classificação de carros por determinada característica e outro pela variação da temperatura em diversas cidades.

A proposta em [2] também é uma extensão do algoritmo fuzzy c-means para o processamento dados intervalares. Simulações são realizadas de um conjunto de dados reais obtidos de um

*Bolsista de Doutorado CAPES

sistema de transporte real. O algoritmo proposto provém do algoritmo fuzzy c-means e permite processar conjunto de dados intervalares e mostra que a proposta desse algoritmo pode ser usado para extrair regras de intervalos fuzzy tipo 2.

Foi proposto em [3] uma outra maneira de trabalhar com dados intervalares. Este é denominado método do centro. Consite em calcular a média aritmética dos valores mínimos e máximos para cada dado intervalar de entrada.

O método proposto em [4] é uma extensão do método do centro [3]. Nessa extensão, os dados são decompostos em dois conjunto de dados. Um consiste nos valores mínimo e o outro consiste nos valores máximos. Atribui-se pesos para essa séries de dados nos valores mínimos e máximos, respectivamente.

Em diversas pesquisas utilizando dados intervalares, como por exemplo descrito em [1][2], são propostas adaptações no algoritmo fuzzy c-means para lidar com dados intervalares. Porém os algoritmo propostos por eles usam graus de pertinências pontuais.

Clusterizando dados de entrada como intervalos, [3] e [4] também não consideram graus de pertinências intervalares.

Os trabalhos [1][2][3] e [4] lidam com dados intervalares mas com uma perspectiva pontual no sentido que os graus de pertinências e as métricas são pontuais.

A extensão do algoritmo fuzzy c-means proposta neste trabalho, não realiza nenhuma conversão dos dados de entrada intervalares para pontuais, nas operações matemáticas do algoritmo, preza-se os dados intervalares. Então a vantagem é que nesta clusterização do algoritmo fuzzy c-means consideram graus de pertinências intervalares propiciando conhecer ainda mais a imprecisão nos dados de entrada. O grande trunfo deste algoritmo é sempre manter os dados de entrada e operações com intervalos e quando necessário calcular a distância de cada ponto ao centro de cada cluster, usar uma métrica intervalar em vez de usar uma métrica pontual como a distância Euclidiana.

A seção 2 mostra as principais operações e funções intervalares especiais na matemática intervalar, destacando-se a forma de calcular a distância entre dois intervalos. Para a seção 3, mostra-se a análise de cluster pontual e a nova extensão do algoritmo proposto baseado no fuzzy c-means. A seção 4 mostra os resultados do algoritmo proposto. Finalmente, a conclusão é discutida na seção 5.

2 Matemática Intervalar

A Matemática Intervalar considera um conjunto de métodos para manipulação de intervalos numéricos que aproximam dados incertos. Na computação científica, os intervalos podem ser aplicados para representar valores desconhecidos e, também para representar valores contínuos. Servem para controlar o erro de arredondamento e para representar dados inexatos, aproximações e erros de truncamento de procedimentos [5]. Estes métodos baseiam-se na definição da Aritmética Intervalar e do produto escalar ótimo [6].

Destacam-se a seguir, as principais operações e funções da matemáticas intervalar utilizadas para este trabalho:

2.1 Operações básicas

Sejam, $X, Y \in \mathbb{IR}$ dois intervalos reais, com $X = [\underline{x}; \overline{x}]$ e $Y = \left[\underline{y}; \overline{y}\right]$. As operações aritméticas intervalares de adição, subtração, multiplicação e divisão são vistas na tabela 1.

Tabela 1: Principais operações sobre intervalos

Descrição	**Operações**
Adição	$X+Y=\left[\left(\underline{x}+\underline{y}\right);\left(\overline{x}+\overline{y}\right)\right]$
Subtração	$X-Y=\left[\left(\underline{x}-\overline{y}\right);\left(\overline{x}-\underline{y}\right)\right]$
Multiplicação	$X\times Y=[min\{\underline{x}\times\underline{y},\underline{x}\times\overline{y},\overline{x}\times\underline{y},\overline{x}\times\overline{y}\};\ max\{\underline{x}\times\underline{y},\underline{x}\times\overline{y},\overline{x}\times\underline{y},\overline{x}\times\overline{y}\}]$
Divisão	$\frac{X}{Y}=\left[\min\left\{\frac{\underline{x}}{\underline{y}},\frac{\underline{x}}{\overline{y}},\frac{\overline{x}}{\underline{y}},\frac{\overline{x}}{\overline{y}}\right\};\max\left\{\frac{\underline{x}}{\underline{y}},\frac{\underline{x}}{\overline{y}},\frac{\overline{x}}{\underline{y}},\frac{\overline{x}}{\overline{y}}\right\}\right]$ com $0\notin\left[\underline{y};\overline{y}\right]$

Os intervalos reais têm várias semânticas como, por exemplo, representação de números reais, quanto mais próximos os extremos do intervalo estiverem do valor "correto", melhor será a representação desse valor. A definição de ordem entre intervalos depende da abordagem ou semântica utilizada. Para este trabalho utilizou-se da ordem de Kulisch-Miranker [7], conforme definição a seguir.

Definição 2.2. *Sejam dois intervalos $X=[\underline{x};\overline{x}]$ e $Y=[\underline{y};\overline{y}]$, a ordem de Kulisch-Miranker define que $[\underline{x};\overline{x}]\leq_K[\underline{y};\overline{y}]\Leftrightarrow\underline{x}\leq\underline{y}$ e $\overline{x}\leq\overline{y}$.*

2.3 Funções Intervalares Especiais

- Define-se os expoentes de uma função intervalar sendo $n\in\mathbb{N}$ conforme mostrado na equação (1).

$$X^n=\begin{cases}[\overline{x}^n;\underline{x}^n] & \text{se } \overline{x}<0 \text{ e } n \text{ for par}\\ [0;\max\left([\underline{x}^n;\overline{x}^n]\right)] & \text{se } \underline{x}<0<\overline{x} \text{ e } n \text{ for par}\\ [\underline{x}^n;\overline{x}^n] & \text{caso contrário}\end{cases} \tag{1}$$

- Define-se a raiz fracionário intervalar na equação (2)

$$\sqrt[n]{X}=\begin{cases}\left[\sqrt[n]{\underline{x}};\sqrt[n]{\overline{x}}\right] & \text{se } \underline{x}\geq 0\\ \uparrow & \text{caso contrário}\end{cases} \tag{2}$$

  Observe que $X^{\frac{m}{n}}=(\sqrt[n]{X})^m$.

2.4 Métrica Intervalar

Em [8] é considerado que a distância entre dois intervalos é um número real, o que não é natural. A para o time B, são 3 pontos, 27 e 30 respectivamente. Pode-se prever essa distância na tabela de pontuação nas próximas duas rodadas. Há como imaginar essa diferença dentre as possiblidades possíveis, porém não há como ter certeza dos resultados de cada jogo. Ao analisar o time A, considerando sua pontuação mínima, não havendo vitórias e a sua pontuação máxima, com duas vitórias, poderíamos representar a previsão da pontuação a ser alcançada na forma de intervalos, sendo $A=[27;33]$ e o outro time, $B=[30;36]$. Com esta representação podemos prever a diferença de pontuação mínima e máxima entre os times após duas rodadas, neste exemplo, o intervalo $[0;9]$.

Foi proposto em [9] uma "métrica"para dados intervalares, na qual a distância entre dois intervalos também é um intervalo, sem perder as características da métrica euclidiana quando se trata de números reais ou intervalos degenerados.

Definição 2.5 (Uma distância essencialmente intervalar)**.** *Sejam X e $Y\in\mathbb{IR}$. A distância essencialmente intervalar entre X e Y, denotada por $d_{MI}(X,Y)$, é o intervalo da equação (3).*

$$d_{MI}(X;Y)=[\min\{d(x;y):x\in X\ e\ y\in Y\};\max\{d(x;y):x\in X\ e\ y\in Y\}] \tag{3}$$

onde $d(x;y)$ é a distância usual ($|x-y|$) entre dois números reais.

Proposição 2.6. *Sejam X e $Y \in \mathbb{IR}$. Então*

$$d_{MI}(X;Y) = \begin{cases} [0; \max(|\overline{X} - \underline{Y}|; |\underline{X} - \overline{Y}|)] & \text{se } X \cap Y \neq \varnothing \\ \left[\min(|\overline{X} - \underline{Y}|; |\underline{X} - \overline{Y}|); \max(|\overline{X} - \underline{Y}|; |\underline{X} - \overline{Y}|)\right] & \text{se } X \cap Y = \varnothing \end{cases} \quad (4)$$

Resultados intervalares carregam consigo a segurança de sua qualidade e o grau de sua incerteza, pois o diâmetro do intervalo solução é um indicativo da influência dos erros dos dados de entrada e dos erros de arredondamento e truncamento no erro do resultado final obtido [10].

3 Análise de *Cluster* Intervalar

De acordo com [11], a técnica de clusterização (clustering) ou agrupamento procura identificar um conjunto de categorias ou classes para descrever os dados.

Segundo [12][13], na clusterização parte-se de uma situação em que não existem classes, somente elementos de um universo. A partir destes elementos, as técnicas de clusterização são responsáveis por definir as classes e enquadrar os elementos.

Entre os diversos algoritmos de clusterização existentes, este trabalho deter-se-á a intervalização do algoritmo proposto por [14] chamado Fuzzy C-Means (FCM).

Baseado na estrutura do FCM [15], é proposto um algoritmo para a clusterização de dados intervalares, denominado Interval Fuzzy C-Means (IFCM).

O IFCM tenta de encontrar conjuntos nos dados minimizando uma função objetiva mostrada na equação (5):

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m} d_{MI}\left(X_i; C_j\right)^2 \quad (5)$$

onde:

- n é o número de dados intervalares;
- c é o número de clusters considerados no algoritmo, o qual deve ser decidido antes da execução;
- m é um fator de fuzziness (um valor maior do que 1) [1];
- X_i é o i-ésimo dado intervalar;
- C_j é o centro (intervalo) do j-ésimo cluster;
- $d_{MI}\left(X_i; C_j\right)$ é a distância intervalar entre X_i e C_j;

A entrada do algoritmo são n dados intervalares, o número de cluster c e o valor m. Suas etapas são:

1. Inicialize μ com subintervalos de $[0;1]$ aleatórios associados a cada par (dados/clusters) tais que para cada par dados/cluster $(X_i; j)$ e $a_j \in \mu_{i,j}$ temos que existem $a_k \in \mu_{i,k}$ para todo $k \in \{1, \ldots, j-1, j+1, \ldots, c\}$ satisfazendo

$$\sum_{k=1}^{c} a_k = 1$$

[1] Só consideramos valores racionais para não complicar o cálculo das equações (5), (6) e (7). Uma vez que na prática são usados m racionais.

2. calcule o centro do cluster j da seguinte maneira:

$$C_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m X_i}{\sum_{i=1}^{n} \mu_{ij}^m} \tag{6}$$

3. calcule um valor inicial (um intervalo de dado) para J usando a equação (5)

4. calcule a tabela de função de pertinência fuzzy intervalar conforme mostrado na equação (7)

$$\mu_{ij} = \frac{\left(\frac{1}{d_{MI}(X_i;C_j)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d_{MI}\left(X_i;C_k\right)}\right)^{\frac{1}{m-1}}} \tag{7}$$

5. retornar a etapa 2 até que uma condição de parada seja alcançada.

Algumas condições de parada possíveis são:

- Um número de iterações pré-fixado foi executado, e pode-se considerar que o algoritmo conseguiu agrupar ("bom o bastante") os dados;
- o usuário informa um valor de parada $\epsilon > 0$, e se

  $$d_{MI}\left(J_U; J_A\right) \leq [\epsilon; \epsilon]$$

  então pára, onde J_A é a função objetiva (equação (5)) calculada da iteração anterior e J_U é a função objetiva da última iteração.

Proposição 3.1. *A cada iteração, cada par dado/cluster* (X_i, j) *e* $a_j \in \mu_{i,j}$ *temos que existem* $a_k \in \mu_{i,k}$ *para todo* $k \in \{1, \ldots, j-1, j+1, \ldots, c\}$ *satisfazendo*

$$\sum_{k=1}^{c} a_k = 1$$

Demonstração. Direto do fato que a versão pontual do cálculo de $\mu_{i,j}$ satisfaz a propriedade de

$$\sum_{k=1}^{c} \mu_{i,k} = 1$$

a cada iteração.

4 Resultados

A implementação do algoritmo foi realizada no ambiente C++ (compilador g++ 4.4), no sistema operacional Linux (Ubuntu 9.04) e utilizou-se a biblioteca C-XSC (versão 2.2).

A entrada dos dados é mostrado na tabela 2. Os dados a serem agrupados é mostrado na coluna X. As colunas Cluster 1 e Cluster 2 são os graus de pertinências (aleatórios) de cada dado de entrada, satisfazendo as condições da etapa 1 do algoritmo.

Tabela 2: Entrada de dados

DADOS	X	CLUSTER 1	CLUSTER 2
1	[1110; 1112]	[0, 04; 0, 05]	[0, 95; 0, 96]
2	[666; 668]	[0, 05; 0, 06]	[0, 94; 0, 95]
3	[100; 102]	[0, 09; 0, 10]	[0, 90; 0, 91]
4	[108; 110]	[0, 10; 0, 11]	[0, 89; 0, 90]
5	[1207; 1209]	[0, 13; 0, 14]	[0, 86; 0, 87]
6	[653; 655]	[0, 70; 0, 71]	[0, 29; 0, 30]
7	[118; 120]	[0, 25; 0, 30]	[0, 70; 0, 75]
8	[198; 200]	[0, 80; 0, 81]	[0, 19; 0, 20]
9	[1195; 199]	[0, 35; 0, 36]	[0, 64; 0, 65]
10	[100; 102]	[0, 84; 0, 85]	[0, 15; 0, 16]

O resultado do algoritmo IFCM é mostrado na tabela 3.

Tabela 3: Resultado do algoritmo IFCM

DADOS	X	CLUSTER 1	CLUSTER 2
1	[1110; 1112]	[0.000230; 0.000257]	[0.999743; 0.999770]
2	[666; 668]	[0.123565; 0.131582]	[0.868418; 0.876435]
3	[100; 102]	[0.999992; 0.999996]	[0.000004; 0.000008]
4	[108; 110]	[0.999996; 0.999999]	[0.000001; 0.000004]
5	[1207; 1209]	[0.001694; 0.001811]	[0.998189; 0.998306]
6	[653; 655]	[0.154389; 0.164010]	[0.835990; 0.845611]
7	[118; 120]	[0.999998; 1.000000]	[0.000000; 0.000002]
8	[198; 200]	[0.999971; 0.999982]	[0.000018; 0.000019]
9	[1195; 199]	[0.001411; 0.001513]	[0.998487; 0.998589]
10	[100; 102]	[0.999992; 0.999996]	[0.000004; 0.000008]

O sistema convergiu após 18 iterações. Os parâmetros de entrada foram 10 dados, 2 clusters, $m = 1, 25$ e $\epsilon = 0, 01$.

5 Conclusões

A análise de cluster não é um processo realizado em apenas uma execução. Em muitas circunstâncias, é necessário uma série de tentativas e repetições. Ainda, não há um critério universal e efetivo para guiar a seleção de atributos e de algoritmos de clusterização. Critérios de validação provêm impressões sobre a qualidade dos clusters, mas como escolher este mesmo critério é ainda um problema que requer mais esforços [16].

Este trabalho apresentou um estudo das principais operações e funções especiais da matemática intervalar e mostrou os procedimentos de análise de clusters. Estudou-se outros algoritmos de clusterização para a entrada de dados intervalares e conclui-se que a proposta neste trabalho apresenta vantagens por nunca realizar uma conversão de dados intervalares para dados pontuais e os graus de pertinências manterem-se como intervalos.

Na extensão intervalar do algoritmo fuzzy c-means proposto neste artigo, houve a aplicação de duas técnicas: a matemática intervalar e a teoria dos conjuntos difusos. Desta forma, é possível tratar os dados de entrada imprecisos em resultados com funções de pertinências intervalares.

Referências

[1] CARVALHO, F. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, USA, v. 28, n. 4, p. 423–437, 2007.

[2] ZHANG, W.; HU, H.; LIU, W. Rules extraction of interval type-2 fuzzy logic system based on fuzzy c-means clustering. *Fuzzy Systems and Knowledge Discovery, Fourth International Conference on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 2, p. 256–260, 2007.

[3] BOCK, H. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2000.

[4] SATO, M.; J., L. *Innovations in Fuzzy Clustering: Theory and Applications (Studies in Fuzziness and Soft Computing)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[5] OLIVEIRA, R.; DIVERIO, T.; D., C. *Fundamentos da Matemática Intervalar*. Instituto de Informática da UFRGS, Porto Alegre, Brasil: Editora Sagra Luzzato, 2001. xi-90 p.

[6] KULISCH, U. *Advanced Arithmetic for the Digital Computer: Design of Arithmetic Units*. Verlag: Softcover, 2002. xii - 141 p.

[7] KULISCH, U.; MIRANKER, W. *Computer Arithmetic in Theory and Practice*. Orlando, FL, USA: Academic Press, Inc., 1981.

[8] MOORE, R. Interval analysis. In: _____. Philadelphia, PA, USA: pub-PH, 1966. p. xi–145.

[9] TRINDADE, R. *Uma Fundamentação Matemática para Processamento Digital de Sinais Intervalares*. Tese (Doutorado) — Universidade Federal de Rio Grande do Norte, Natal, Brasil, 2009.

[10] DIMURO, G. *Domínios Intervalares da Matemática Computacional*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 1991.

[11] FAYYAD, U. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. [S.l.]: AAAI/MIT Press, 1996.

[12] HAN, J. et al. Intelligent query answering by knowledge discovery techniques. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, p. 373–390, 1996.

[13] AGRAWAL, R. Data mining: The quest perspective. *Australian Computer Science Comm. — Proc. 7th Australasian Database Conf., ADC*, v. 18, n. 2, p. 119–120, 1996.

[14] BEZDEK, J. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[15] COX, E. *Fuzzy Modeling and Genetic Algorithms For Data Mining and Exploration*. San Francisco: Morgan Kaufmann, 2005. Elsevier.

[16] CAVALCANTI, N. J. *Clusterização Baseada em Algoritmos Fuzzy*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, Brasil, 2006.