Metodologia Incremental para Agrupamento em Fluxos Contínuos de Dados

José Maia Neto[1], Cristiano Leite de Castro[2],
André Paim Lemos[3], and Liliane dos Reis Gade[4]

Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, 31270-901
Belo Horizonte, MG, Brasil
[1] jmnt@ufmg.br,[2] crislcastro@ufmg.br,[3] andrepaim@ufmg.br[4]
liligade4@gmail.com

Resumo Este artigo apresenta uma metodologia para agrupamento incremental de dados em fluxos contínuos. O método proposto se baseia nos conceitos de tipicidade e excentricidade e no algoritmo CEDAS, recentemente introduzidos. A cada nova amostra recebida, atualiza-se uma estrutura de micro grupos os quais armazenam, dentre outros parâmetros, a densidade local dos dados e a tipicidade local. Em seguida, uma estrutura de macro grupos é atualizada como sendo uma soma das tipicidades dos micro grupos que se sobrepõem ponderadas pela densidade local de cada um destes micro grupos. Ao final tem-se um modelo de mistura de densidades locais que possui a capacidade de agrupar dados de distribuições arbitrárias e gerar como saída um valor de pertinência de uma amostra para cada agrupamento. Os resultados preliminares, com bases de dados sintéticas, mostraram que o algoritmo proposto é promissor para aplicações de agrupamento online.

Palavras-chave: Tipicidade, Agrupamento Incremental, Densidade, Fluxos de Dados.

1 Introdução

A necessidade de se extrair conhecimentos de fluxos de dados contínuos tem motivado o interesse no desenvolvimento de algoritmos de agrupamento online. Entretanto, a aplicação destes algoritmos em sistemas reais ainda apresenta grandes desafios. Frequentemente, os fluxos de dados provêm de sistemas não estacionários, dos quais se tem pouca informação *a priori*. Este fator dificulta a escolha de parâmetros como a função de distribuição geradora dos dados e a quantidade de agrupamentos. Para lidar com esses problemas, muitos trabalhos nesta área se baseiam em janelas deslizantes [14], ou estratégias híbridas *online/offline* [11]. Em geral, tais métodos apresentam bons resultados e um custo computacional aceitável para grande parte das aplicações. O seu desempenho porém, é dependente de, dentre outros parâmetros, a escolha do tamanho da janela, o que

nem sempre é uma tarefa simples. Outros trabalhos apresentam metodologias de agrupamento puramente incrementais [6]. Tais metodologias utilizam apenas a amostra atual para fazer os cálculos e armazenar as informações descritivas de cada agrupamento (centro, raio, etc.) em um protótipo o qual representa todas as amostras pertencentes a um determinado grupo. Uma vantagem destes algoritmos é que são capazes de gerar novos agrupamentos de acordo com o comportamento dos dados ao longo do tempo. No entanto, apresentam dificuldades ao lidar com dados provenientes de diferentes distribuições.

O presente trabalho apresenta uma proposta de algoritmo incremental para agrupamento de fluxos de dados. O algoritmo proposto consiste de duas etapas. Na primeira, a cada nova amostra recebida, uma estrutura de micro grupos é atualizada. Em seguida, a estrutura de micro grupos é utilizada para gerar macro grupos como sendo combinações dos micro grupos que possuem sobreposição. Após a atualização das estruturas, a amostra corrente é atribuída ao grupo cuja soma de tipicidades [3] dos micro grupos ponderadas pela sua densidade local apresenta maior valor. O método proposto apresenta vantagens em relação a algoritmos do estado-da-arte ao passo que pode agrupar dados em formas ou distribuições arbitrárias além de retornar como saída um valor de pertinência de uma amostra com relação a cada grupo, o que o credencia para aplicações que requerem *soft clustering*.

Este artigo está organizado da seguinte forma. Na seção 2 é apresentada uma revisão dos métodos de agrupamento *online*, bem como a estrutura da metodologia proposta. Na seção 3 a abordagem proposta é descrita. Na seção 4 os resultados dos experimentos são apresentados e discutidos. Finalmente na seção 5 são apresentadas as conclusões e os trabalhos futuros.

2 Revisão da Literatura

O método proposto neste trabalho baseia-se nos conceitos de excentricidade e tipicidade (TEDA - *Typicallity and Eccentricity Data Analysis*) introduzidos por [3]. Esta abordagem consiste em um algoritmo para detecção de anomalias que se caracteriza por modelar uma distribuição não paramétrica dos dados baseado apenas na proximidade de uma amostra particular para todo um conjunto de dados. As equações do TEDA podem ser calculadas recursivamente para algumas métricas de distância, tais como Euclidiana, Mahalanobis e Cosseno, sendo factível para aplicações online.

2.1 TEDA

O TEDA se baseia no conceito de proximidade acumulada. Dado um vetor de entradas $\vec{x} = \{x_1, x_2, \ldots, x_d\} \in \mathbb{R}^d$ no instante k, a proximidade acumulada é calculada como [3].

$$\pi_k(\vec{x}) = \sum_{i=1}^{k} d(\vec{x}, \vec{x}_i) \ , \tag{1}$$

em que $d(a, b)$ representa uma função de distância entre dois pontos a e b, e k representa o instante de tempo em que um determinado dado de entrada $\vec{x}_k$ é amostrado.

A partir de $\pi_k(\vec{x})$, obtém-se a excentricidade, que é uma medida de dissimilaridade de uma amostra de entrada $\vec{x}$(o quão excêntrica) em relação aos demais dados de entrada amostrados até o instante k. A excentricidade é definida como.

$$\xi_k(\vec{x}) = 2\frac{\pi_k(\vec{x})}{\sum_{i=1}^{k} \pi_k(\vec{x}_i)} \ , \ \sum_{i=1}^{k} \pi_k(\vec{x}_i) > 0 \ , \ k \geq 2. \tag{2}$$

A tipicidade, por sua vez, representa o quão típica é uma amostra com relação às demais amostras recebidas até o instante k, sendo calculada como o dual da excentricidade.

$$\tau_k(\vec{x}) = 1 - \xi_k(\vec{x}) \ , \ \pi_k(\vec{x}_i) > 0 \ , \ k \geq 2 \tag{3}$$

Para a distância Euclidiana, tanto a tipicidade quanto a excentricidade podem ser calculados recursivamente [3].

$$\xi_k(\vec{x}) = \frac{1}{k} + \frac{(\vec{\mu}_x^k - \vec{x})^T(\vec{\mu}_x^k - \vec{x})}{k(\sigma_x^k)^2}, \tag{4}$$

em que $\vec{\mu}_x^k$ é a média e $(\sigma_x^k)^2$ a variância, que também são calculados recursivamente.

$$\vec{\mu}_x^k = \frac{k-1}{k}\vec{\mu}_x^{k-1} + \frac{\vec{x}_k}{k} \ , \ k \geq 1 \ , \ \vec{\mu}_x^1 = \vec{x}_1. \tag{5}$$

$$(\sigma_x^k)^2 = \mu_{\vec{x}^T\vec{x}}^k - (\vec{\mu}_x^k)^T\vec{\mu}_x^k \ , \ (\sigma_x^1)^2 = 0 \tag{6}$$

$$\mu_{\vec{x}^T\vec{x}}^k = \frac{k-1}{k}\mu_{\vec{x}^T\vec{x}}^{k-1} + \frac{(\vec{x}_k)^T\vec{x}_k}{k} \ , \ k \geq 1 \ , \ \mu_{\vec{x}^T\vec{x}}^1 = \vec{x}_1^T\vec{x}_1 \tag{7}$$

O conceito de tipicidade e excentricidade é ilustrado na Fig. 1[3]. O ponto A está mais distante do conjunto de dados (pontos pretos) e, portanto, apresenta maior excentricidade e menor tipicidade do que o ponto B.

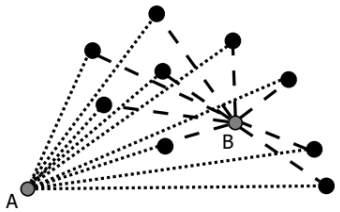


Figura 1. Conceitos do TEDA [5].

A versão normalizada da excentricidade é definida pela Eq. 8 [3]. A excentricidade normalizada $\zeta_k(\vec{x})$ pode ser utilizada na definição de um limiar, baseado na desigualdade de Chebyshev, para detecção de *outlier* [2][13]. De modo geral, a condição expressa pela Eq. 9 define, com base na excentricidade normalizada e na desigualdade de Chebyshev, se uma amostra está "$m\sigma$" distante da média, onde m é um parâmetro que define "quantos desvios-padrão" distante da média uma amostra deve estar para que seja considerada um *outlier*.

$$\zeta_k(\vec{x}) = \frac{\xi_k(\vec{x})}{2} \tag{8}$$

$$\zeta_k(\vec{x}) > \frac{m^2+1}{2k},\ m > 0 \tag{9}$$

Devido à rapidez e facilidade de implementação, extensões do TEDA têm sido propostas para lidar com problemas de classificação [9] , regressão [8] e agrupamento [6] em fluxos contínuos de dados. Uma característica importante deste modelo é que a tipicidade se assemelha (embora seja diferente) à uma função massa de probabilidade [4], sendo útil para aplicações onde se faz necessário o uso de *soft clustering* ou modelar a distribuição dos dados quando esta é desconhecida *a priori*.

2.2 Algoritmos de Agrupamento Incremental

Em [14] é apresentada uma revisão sobre métodos de agrupamento *online*. Os algoritmos apresentados no artigo utilizam, em sua maioria, janelas deslizantes para detectar novos grupos. Em resumo, esses algoritmos atualizam as informações dos grupos com base apenas nas amostras contidas em uma janela de dados, esta janela por sua vez contém os dados mais recentes, ou mais "relevantes", onde o conceito de relevância varia para cada método. O tamanho da janela é definido de acordo com os recursos computacionais disponíveis. Outras metodologias, como a versão incremental do algoritmo *k-means* [12] não utilizam janelas de dados, entretanto necessitam que a quantidade de grupos seja conhecida *a priori*. Recentemente, alguns trabalhos apresentaram modelos de agrupamento incremental baseados no conjunto de dados [5] [10]. Estes modelos apresentam a vantagem de necessitar de pouco conhecimento prévio sobre os dados devido a sua capacidade de evoluir tanto a estrutura (quantidade de grupos) quanto os parâmetros (centros , raios, etc.) do modelo ao longo do tempo, a cada nova amostra processada .

Em geral, os algoritmos de agrupamento *online*, partem de algumas premissas: a) os dados possuem uma distribuição fixa b) necessitam que se conheça a quantidade de agrupamentos a priori. Por sua vez, os métodos que se propõem a definir a quantidade de grupos de maneira automática necessitam que os dados sejam amostrados de maneira ordenada (ordenação por grupo). Para muitos problemas práticos em que os dados de entrada são amostrados na forma de fluxos contínuos, estas propriedades podem não ser válidas. Assim, o algoritmo de agrupamento deve ser robusto à elas de modo a generalizar a solução

para a maior gama de aplicações, apresentando um procedimento automatizado e orientado aos próprios dados.

Estudo recente apresentado em [1] propõe uma formalização para o problema de agrupamento incremental de dados e aponta algumas dificuldades encontradas pelos métodos da literatura. O trabalho sugere que para um algoritmo incremental conseguir agrupar dados em grupos consistentes (grupos com mínima distância intra grupo e máxima distância entre grupos) é necessário dividir o problema em micro grupos. O trabalho apresentado por [7] apresentou um método de agrupamento incremental com estas características, o algoritmo CEDAS (*Clustering Evolving Data Streams*). Este método divide o problema em micro grupos e depois constrói um grafo com os centros dos micro grupos que possuem sobreposição. Ao final do processo, cada grafo formado corresponde a um macro grupo. Embora este método possua a capacidade de agrupar dados de distribuições e formas arbitrárias, não é possível se obter informações sobre a densidade local dos grupos sendo possível apenas saber se uma amostra pertence ou não pertence à cada grupo.

O presente trabalho apresenta uma nova abordagem para o agrupamento incremental. A cada nova amostra recebida, o método proposto atualiza uma estrutura de micro grupos e uma estrutura de macro grupos. A estrutura de micro grupos consiste em grupos com raio limitado a um limiar cujo valor varia para cada aplicação. A estrutura de macro grupos consiste em um modelo de soma ponderada de tipicidades, calculado como uma combinação de micro grupos que estão próximos. Ao final tem-se como resultado grupos com diferentes formas e distribuições. A abordagem proposta é detalhada a seção seguinte.

3 Abordagem proposta

A abordagem proposta é realizada em duas etapas, de forma similar a [7]. Na primeira etapa, a cada nova amostra x_k recebida no instante de tempo k, calcula-se a distância de x_k para o centro μ_k de todos os micro grupos existentes e seleciona-se o micro grupo mais próximo. Caso x_k esteja dentro do raio de influência do micro grupo mais próximo (Eq. 10), considerando-se o grupo i como sendo o mais próximo de x_k, os parâmetros de média e variância, excentricidade e tipicidade $(1 - \xi_k(\vec{x}_k)^i)$ deste micro grupo são atualizados recursivamente.

$$dist(x_k, \mu_k^i) < r_0 \tag{10}$$

$$S_k^i = S_{k-1}^i + 1 \tag{11}$$

$$\xi_k(\vec{x}_k)^i = \frac{1}{S_k^i} + \frac{(\vec{\mu}_k^i - \vec{x}_k)^T(\vec{\mu}_k^i - \vec{x}_k)}{k(\sigma_k^i)^2}. \tag{12}$$

$$\vec{\mu}_k^i = \frac{S_k^i - 1}{S_k^i}\vec{\mu}_{k-1}^i + \frac{\vec{x}_k}{S_k^i} \tag{13}$$

$$\mu^i_{(\vec{x}^T\vec{x})_k} = \frac{S^i_k - 1}{S^i_k}\mu^i_{(\vec{x}^T\vec{x})_{k-1}} + \frac{(\vec{x}_k)^T\vec{x}_k}{S^i_k} \tag{14}$$

$$(\sigma^i_k)^2 = \mu_{(\vec{x}^T\vec{x})_k} - (\vec{\mu}^i_k)^T\vec{\mu}^i_k \tag{15}$$

Em que S^i_k correnponde à quantidade de amostras pertencentes ao grupo i no instante k e r_0 é o um parâmetro defindo pelo usuário que limita o raio de influência de um micro grupo sobre as amostras. Caso a condição imposta pela Eq. 10 não seja verdadeira para nenhum micro grupo existente, então um novo micro grupo é criado com os seguintes parâmetros.

$$S^{new}_k = 1 \tag{16}$$

$$\vec{\mu}^{new}_k = \vec{x}_k. \tag{17}$$

$$\mu^{new}_{(\vec{x}^T\vec{x})_k} = \vec{x}^T_k\vec{x}_k \tag{18}$$

$$(\sigma^{new}_k)^2 = 0 \tag{19}$$

As etapas de atualização da estrutura dos micro grupos são descritas em detalhes pelo Algoritmo 1. Os micro grupos encontrados para uma base de dados sintética com $r_0 = 0.03$ são ilustrados na Figura 2b. O valor de r_0 tem a função de limitar o raio dos micro grupos, dessa forma, maiores valores resultarão menores quantiades de micro grupos criados ao preço de se perder, eventualmente a qualidade dos grupos formados. Com base nos experimentos deste artigo, sugere-se valores de r_0 entre 0.02 e 0.05 para bases de dados normalizadas no hipercubo unitário.

Em seguida a estrutura de macro grupos é atualizada. A Figura 2 ilustra este procedimento de atualização para uma base de dados sintética. De posse do conjunto de micro grupos atual (Figura 2b), o algoritmo encontra todos os conjuntos de micro grupos que apresentam sobreposição (Figura 2c), de forma semelhante à estrutura de grafos mostrada em [7]. Para isso, calcula-se a distância do centro do micro grupo (mg) mais próximo de x_k para todos os centros dos outros micro grupos, caso esta distância seja menor que r_0 então significa que eles se sobrepõem, desta forma mg é adicionado ao macro grupo referente ao macro grupo ao qual ele apresenta sobreposição.

Em seguida a tipicidade de x_k para cada macro grupo é calculada como sendo a soma das tipicidades dos micro grupos que o constitui, ponderados pela respectiva densidade local w. Seja M o conjunto dos micro grupos pertencentes ao macro grupo n, a tipicidade T_n deste macro grupo é calculado como:

$$T_n = \sum_{j \in M} w^j_k \tau_k(\vec{x}_k)^j \tag{20}$$

Algoritmo 1: Atualização dos Micro Grupos

Entrada: x_k, r_0
Saída: *micro grupos*
início
  while *novas amostras disponíveis* **do**
    if $k = 1$ **then**
      %Cria o primeiro micro grupo;
      $nmicrogrupos = 1$;
      $S_1^1 = 1$;
      $\vec{\mu}_1^1 = \vec{x}_1$;
      $\mu^1_{(\vec{x}^T\vec{x})_1} = \vec{x}_1^T\vec{x}_1$;
      $(\sigma_1^1)^2 = 0$;
    else
      %Encontra o micro grupo mais próximo
       $minDist = argmin(dist(x_k, \mu_k^i))$, $i = 1, 2, \ldots, nmicrogrupos$;
      if $minDist < r_0$ ***e*** $S_k^{minDist} \geq 2$ **then**
        Adiciona x_k *ao micro grupo mais próximo*;
        Atualiza $S_k^{minDist}$ *conforme Eq. 11*;
        Atualiza $\vec{\mu}_k^{minDist}$ *conforme Eq. 13*;
        Atualiza $\mu^{minDist}_{(\vec{x}^T\vec{x})_k}$ *conforme Eq. 14*;
        Atualiza $(\sigma_k^{minDist})^2$ *conforme Eq. 15*;
      else
        %Cria um novo micro grupo;
        $nmicrogrupos = nmicrogrupos + 1$;
        $S_k^{new} = 1$;
        $\vec{\mu}_k^{new} = x_k$;
        $\mu^{new}_{(\vec{x}^T\vec{x})_k} = \vec{x}_k^T\vec{x}_k$;
        $(\sigma_k^{new})^2 = 0$;
      end
    end
  end
fim

$$w_k^j = \frac{D_k^j}{\sum\limits_{j \in M} D_k^j} \tag{21}$$

$$D_k^j = \frac{1}{\zeta_k(\vec{x}_k)^j} \tag{22}$$

Ao final de cada passo, a estrutura de macro grupos representa a solução do agrupamento dos dados. Devido aos micro grupos serem combinados de maneira similar a uma mistura de densidades tem-se, uma estimativa de densidade de forma arbitrária para cada grupo, sendo possível se calcular qual a pertinência de uma nova amostra para cada macro grupo. Para avaliar a qual macro grupo

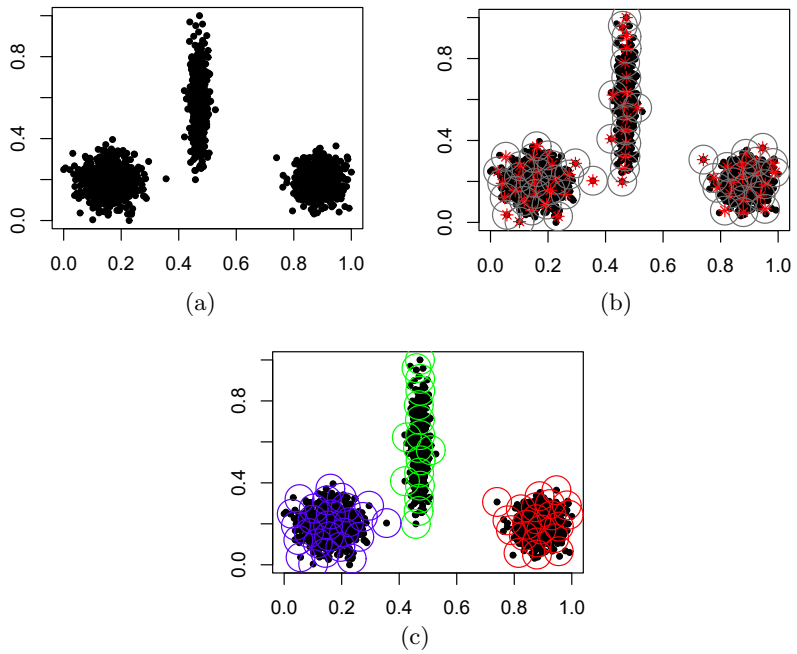


Figura 2. (a) Base de Dados; (b) Identificação dos micro grupos (c) identificação dos macro grupos

uma amostra pertence é calculada a tipicidade de cada macro grupo T_n e a amostra é atribuída ao macro grupo de maior valor.

$$grupo(x_k) = argmax(T_n)\ ,\ n = 1, 2, \ldots, nmacrogrupos \tag{23}$$

O procedimento de atualização dos macro grupos é detalhada no Algoritmo 2.

4 Experimentos e Resultados

Os experimentos foram executados utilizando-se o software R. Para avaliar o método proposto foram realizados testes com bases de dados sintéticas do pacote *mlbench*. O resultado final das densidades estimadas para algumas bases de dados podem ser vistos na Figura 3.

A Figura 3 mostra os resultados obtidos após o agrupamento incremental dos dados utilizando o algoritmo proposto. Verifica-se que o algoritmo apresenta a capacidade de modelar cada grupo através de uma região de densidades de forma arbitrária, baseado na proximidade dos dados. É possível verificar que dentro de um mesmo grupo podem existir lacunas com menor probabilidade de ocorrência de amostras. É interessante ressaltar que, devido a natureza adaptativa do

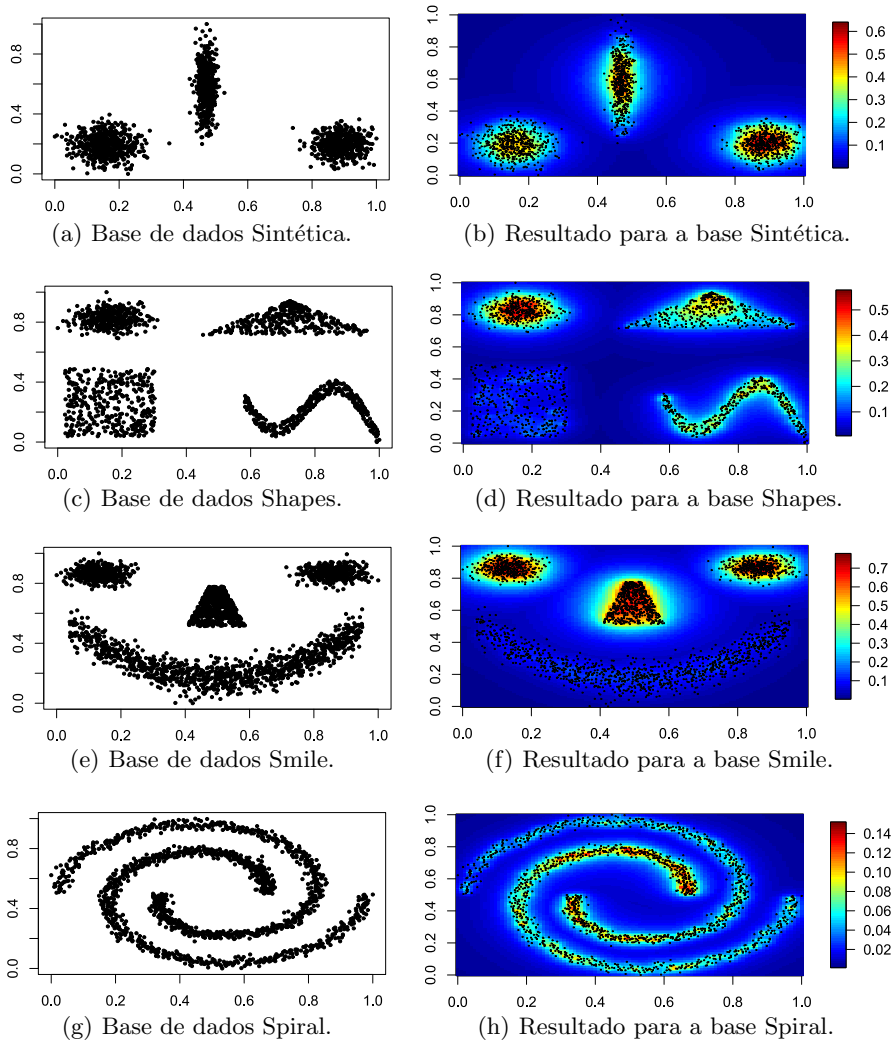
Algoritmo 2: Atualização dos Macro Grupos

Entrada: x_k, r_0, *micro grupos*
Saída: tipicidade de x_k para cada macro grupo
início
  while *novas amostras disponíveis* **do**
    mg = *micro grupo mais prximo de* x_k;
    for i = *1:nmicroclusters* **do**
      if $dist(\mu_k^{mg}, \mu_k^i) > r_0$ **then**
        $flag = 1$;
      else
        $flag = 0$;
      end
    end
    if $flag == 1$ **then**
      Adiciona mg ao macro grupo ao qual ele apresenta sobreposição;
    else
      Cria um novo macro grupo com os mesmos parâmetros de mg;
    end
    Calcula T_n de acordo com Eq. 20 , $n = 1, 2, \ldots nmacrogrupos$;
    Atribui x_k ao grupo que apresenta maior valor de T_n;
  end
fim

algoritmo, conforme novos dados vão sendo amostrados nestes locais de baixa densidade, os parâmetros são atualizados de modo a aumentar a densidade local daquela região do espaço, aumentando assim a zona de atração para os pontos que forem amostrados próximos à ela nos instantes de tempo seguintes.

A qualidade dos agrupamentos foi avalida de acordo com o *recall*, *precision*(métricas externas) e *Silhouette*(métrica interna). Os resultados obtidos para estas métricas, bem como as características das bases de dados utilizadas nos experimentos são apresentados na Tabela 1. Os resultados mostram que, em geral, o algoritmo proposto foi capaz de agrupar os dados de maneira satisfatória para as bases *smile, shapes, sintética e spirals*, comprovando a sua capacidade de agrupar dados de diferentes formas e distribuições. O resultado para a base de dados *s1*, porém, foi insatisfatório, em relação às demais bases. O algoritmo proposto apresentou dificuldades ao lidar com os dados desta base, devido os mesmos apresentarem um maior grau de sobreposição entre os grupos e similaridade entre as formas. Uma solução para este problema seria o emprego de uma metodologia para fusão e separação automática dos macro grupos.

De maneira geral, os testes mostraram que o algoritmo apresenta a performance satisfatória no agrupamento e modelagem de densidade de fluxos de dados apresentando agrupamentos com boa coesão e densidades locais com boa representatividade das características inerentes à cada base de dados.



(a) Base de dados Sintética. (b) Resultado para a base Sintética.

(c) Base de dados Shapes. (d) Resultado para a base Shapes.

(e) Base de dados Smile. (f) Resultado para a base Smile.

(g) Base de dados Spiral. (h) Resultado para a base Spiral.

Figura 3. Resultados obtidos.

5 Conclusões

O presente artigo apresentou uma metodologia para agrupamento incremental de fluxos de dados a partir de um modelo mistura de densidades locais. A abordagem proposta é baseada na proximidade dos dados e apresentou capacidade de encontrar agrupamentos coesos e de formas arbitrárias o que facilita a sua implementação em problemas nos quais não se conhece a distribuição dos dados. O fato de ser recursivo e atualizar os seus parâmetros e a sua estrutura a cada nova amostra processada o credencia para aplicações de aprendizado *online* dando-lhe a característica de se adaptar a mudanças de conceito nos dados ao

Tabela 1. Resultados Experimentais

Base de dados	*Amostras*	*Grupos*	*Grupos encontrados*	*Recall*	*Precision*	*Silhouette*
smile	2500	4	4	0.97	0.98	0.64
shapes	1500	4	4	1.00	1.00	0.70
sintética	1500	3	3	1.00	1.00	0.76
spirals	1500	2	2	1.00	1.00	0.035
s1	5000	15	5	0.951	0.22	0.46

longo do tempo. Além disso, a saída do algoritmo representa a "pertinência" de uma amostra para cada grupo, o que facilita sua aplicação à problemas de *soft clustering*. Os resultados também mostraram que alguns aspectos precisam ser trabalhados. Um deles se refere à criação de agrupamentos extras e/ou agrupamentos pouco coesos degradou o desempenho do agrupamento final para bases de dados com maior sobreposição de amostras de diferentes grupos. Desse modo, pretende-se como trabalho futuro aplicar técnicas de fusão e separação de grupos para aumentar a acurácia do agrupamento e implementar metodologias para identificar e desconsiderar possíveis *outliers* na atualização do modelo para que o método seja generalizado para problemas mais complexos e bases de dados reais.

Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da FAPEMIG, CNPQ e CAPES - Brasil.

Referências

1. Ackerman, M., Dasgupta, S.: Incremental clustering: The case for extra clusters. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 307–315. Curran Associates, Inc. (2014), `http://papers.nips.cc/paper/5608-incremental-clustering-the-case-for-extra-clusters.pdf`
2. Angelov, P.: Anomaly detection based on eccentricity analysis. In: 2014 IEEE Symposium on Evolving and Autonomous Learning Systems (EALS). pp. 1–8 (Dec 2014)
3. Angelov, P.: Outside the box: an alternative data analytics framework. Journal of Automation Mobile Robotics and Intelligent Systems 8(2), 29–35 (2014)
4. Angelov, P., Gu, X., Kangin, D.: Empirical data analytics. International Journal of Intelligent Systems (2017)
5. Bezerra, C.G., Costa, B.S.J., Guedes, L.A., Angelov, P.P.: A new evolving clustering algorithm for online data streams. In: Evolving and Adaptive Intelligent Systems (EAIS), 2016 IEEE Conference on. pp. 162–168. IEEE (2016)
6. Costa, B.S.J., Bezerra, C.G., Guedes, L.A., Angelov, P.P.: Unsupervised classification of data streams based on typicality and eccentricity data analytics. In: 2016

IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 58–63 (July 2016)
7. Hyde, R., Angelov, P., MacKenzie, A.: Fully online clustering of evolving data streams into arbitrarily shaped clusters. Information Sciences 382, 96–114 (2017)
8. Kangin, D., Angelov, P.: Evolving clustering, classification and regression with teda. In: Neural Networks (IJCNN), 2015 International Joint Conference on. pp. 1–8. IEEE (2015)
9. Kangin, D., Angelov, P., Iglesias, J.A.: Autonomously evolving classifier tedaclass. Information Sciences 366, 1 – 11 (2016), http://www.sciencedirect.com/science/article/pii/S002002551630336X
10. Lemos, A., Gomide, F., Caminhas, W.: Multivariable gaussian evolving fuzzy modeling system. Fuzzy Systems, IEEE Transactions on 19(1), 91–104 (2011)
11. Masud, M., Gao, J., Khan, L., Han, J., Thuraisingham, B.M.: Classification and novel class detection in concept-drifting data streams under time constraints. IEEE Transactions on Knowledge and Data Engineering 23(6), 859–874 (June 2011)
12. Pham, D.T., Dimov, S.S., Nguyen, C.D.: An incremental k-means algorithm. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 218(7), 783–795 (2004), http://dx.doi.org/10.1243/0954406041319509
13. Saw, J.G., Yang, M.C.K., Mo, T.C.: Chebyshev inequality with estimated mean and variance. The American Statistician 38(2), 130–132 (1984), http://www.jstor.org/stable/2683249
14. Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., Carvalho, A.C.P.L.F.d., Gama, J.a.: Data stream clustering: A survey. ACM Comput. Surv. 46(1), 13:1–13:31 (Jul 2013), http://doi.acm.org/10.1145/2522968.2522981