Uso de Índices de Validade de Agrupamento para Ajuste de Protótipos

Lucas Matheus de Moraes Florentino
Faculdade de Ciências Exatas e Tecnologia
Universidade Federal da Grande Dourados
Dourados-MS, Brasil
lucas.florentino2014@gmail.com

***Resumo*—Geralmente, algoritmos de agrupamento particionais fazem uso de protótipos representantes de grupos, tornando eficiente a análise dos dados. Tais protótipos devem se tornar centroides dos respectivos grupos para generalizar as características dos mesmos, o que nem sempre ocorre, comprometendo então a confiabilidade da análise dos resultados. Assim, este artigo propõe avaliar a representatividade de protótipos aos seus grupos por meio de índices de validade interna e, ajustar a posição dos mesmos, se necessário. Experimentos foram realizados com o algoritmo k-Médias para três índices de validade de grupos: Soma do Erro Quadrático, Silhueta Simplificada e Xie-Beni. Os resultados mostram a eficácia da proposta, com ênfase para o índice da Soma do Erro Quadrático.**

***Palavras-chave*— Agrupamento de dados, índices internos de validade, k-Médias.**

I. INTRODUÇÃO

A organização de elementos com características similares entre si, em grupos, é tarefa essencial para discriminá-los e compreender como se relacionam entre si. Esta tarefa de segmentação recebe diferentes nomes de acordo com a área de estudo em que é aplicada. Em Mineração de Dados é conhecida como Agrupamento de Dados, enquanto em Aprendizado de Máquina é chamada de Classificação Não-Supervisionada, e Taxonomia em Biologia para a classificação dos seres vivos. Elementos a serem agrupados podem ser referenciados por diferentes nomenclaturas também, tais como dados ou objetos, padrões, vetores-característica. Agrupamento é uma tarefa que possui diversas aplicações práticas, tais como Bioinformática (Masood e Khan, 2015), Negócios (Popkova et al., 2017), Engenharia estrutural (Saiful Bahari et al., 2017), Processamento de imagens (Huang et al., 2019; Jain et al., 1988), Sumarização (Alguliyev et al., 2019).

Por se tratar de uma tarefa não-supervisionada, isto é, os rótulos dos elementos a serem agrupados não são conhecidos *a priori*, alguma medida de (dis)similaridade entre eles deve ser adotada. Geralmente, uma medida de distância é aplicada, uma vez que os dados são representados por vetores em um espaço $\mathbb{R}^d$, tal que d é a dimensão (número de características) dos dados. Dessa maneira, a tarefa objetiva minimizar distâncias intragrupo e maximizar distâncias intergrupo (Han e Kamber, 2001), isto é, produzir grupos coesos e separados entre si.

Formalmente, $X = \{x_1, x_2, \dots, x_n\}$ é o conjunto de dados a ser agrupado em k grupos disjuntos $C = \{c_1, c_2, \dots, c_k\}$, com $n > k$ (Hruschka e Ebecken, 2003):

$$C_1 \cup C_2 \cup \dots \cup C_k = X \quad (1)$$

$$C_i \neq \emptyset, \forall i, 1 \leq i \leq k \quad (2)$$

$$C_i \cap C_j = \emptyset, \forall i \neq j, 1 \leq i \leq k, 1 \leq j \leq k \quad (3)$$

Algoritmos de agrupamento que particionam objetos em grupos fazem parte de uma classe de algoritmos chamada Particional, em que o número de grupos deve ser informado como parâmetro de entrada do algoritmo, ou estimado dinamicamente ao longo das iterações (de Castro e Timmis, 2002; Szabo e de França, 2015). Tais algoritmos fazem uso de vetores representantes de grupos, denominados protótipos. Um protótipo é um vetor no espaço $\mathbb{R}^d$, geralmente inicializado aleatoriamente pelo algoritmo de agrupamento, mas atualizado iterativamente. Espera-se que ele se torne um centroide representante de um grupo, ou que permaneça em região de maior densidade de objetos em um grupo, de modo que possa generalizar suas características. O uso de protótipos possibilita reduzir o tamanho da base de dados a ser agrupada, facilitando a análise dos mesmos.

Entretanto o protótipo pode ficar posicionado na fronteira entre grupos e não se tornar um centroide, assim o protótipo não apresenta boa representatividade para um respectivo grupo, isto compromete a confiabilidade da análise do resultado. Assim, (Szabo e Ruckl, 2021) propuseram um método de ajuste de posição de protótipos, o qual avalia a qualidade dos grupos e ajusta a posição dos respectivos protótipos, se necessário. A qualidade de cada grupo é avaliada por uma métrica que considera distâncias intragrupo (coesão do grupo), e neste artigo o Método é avaliado também para duas outras métricas, as quais consideram, simultaneamente, distâncias intragrupo e intergrupo (separação entre grupos). Este artigo está organizado da seguinte maneira: a Seção II apresenta um dos algoritmos clássicos da literatura de agrupamento de dados (*k*-Médias) e na Seção III são apresentadas três métricas de avaliação de agrupamento. O Método de Ajuste de Protótipos é descrito na Seção IV e os resultados são avaliados e discutidos na Seção V para três tipos de análises. O artigo é concluído na Seção VI com a proposta de trabalhos futuros.

II. ALGORITMO K-MÉDIAS

O algoritmo *k*-Médias, proposto em 1967 (MacQueen), é um dos algoritmos da literatura de agrupamento particional de dados. Primeiramente, ele seleciona *k* objetos da base de dados a ser agrupada, aleatoriamente, de modo a tornarem-se protótipos representantes de grupos. O algoritmo possui dois passos, basicamente: (i) atualizar os protótipos e, (ii) atualizar a matriz de pertinências ($\mathbf{U}|\mathbf{U} = \{\mu_{11}, \mu_{12}, \dots, \mu_{ij}\}, \mu_{ij} \in \{0,1\}$), a qual representa o grau de associação entre objeto de índice *i* e grupo de índice *j*. Assim, $\mu_{ij} = 1$ se o objeto de índice *i* pertence ao grupo de índice *j*, ou $\mu_{ij} = 0$, caso contrário. Além da matriz **U**, os protótipos são atualizados iterativamente:

$$\mathbf{c}_j = \frac{\sum_{i=1}^{n} \mu_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \mu_{ij}} \quad (4)$$

O *k*-Médias objetiva minimizar a seguinte função:

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathbf{c}_j} ||\mathbf{x} - \mathrm{c}_j||^2 \quad (5)$$

Em que *SSE* é a Soma do Erro Quadrático da partição; **x** é o objeto pertencente ao grupo c_j, e $||.||$ é a norma Euclidiana.

III. Indices Internos de Validade de Agrupamento

A tarefa de verificar se uma partição está correta é desafiadora, uma vez que não há um modelo a ser comparado em problemas reais. A tarefa de avaliar a qualidade de uma partição é conhecida como validade de agrupamento, em que índices de validade são utilizados para esta finalidade (Jain et al., 1988).

Os métodos de validade presentes na literatura são classificados em dois grupos, principalmente: (i) validação externa e (ii) validação interna (Sameh et al., 2009). A validação externa é utilizada para avaliar a qualidade dos grupos a partir de seus rótulos previamente conhecidos; desta maneira, são aplicados em problemas de *benchmarking*. A validação interna, por sua vez, utiliza como parâmetros de avaliação valores do próprio conjunto de dados, isto é, distâncias intragrupo e/ou intergrupo. Assim, diversos índices de validação foram propostos, tais como Xie-Beni (Xie e Beni, 1991), Silhueta (Rousseeu, 1987), Silhueta Simplificada (Vendramin et al., 2010), dentre outros.

Índices internos podem ser aplicados para estimar o número ótimo de grupos da base de dados a ser agrupada, ou para guiar o algoritmo de agrupamento (como um critério de parada do mesmo), ou ainda para a validação da partição por ele produzida (Bakhtiar et al., 2016). A seguir são apresentados três dos índices de validade de grupos utilizados neste trabalho: índice da Soma do Erro Quadrático, índice Silhueta Simplificada e índice Xie-Beni.

A. Índice da Soma do Erro Quadrático

A Soma do Erro Quadrático é uma medida que avalia a distância intragrupo [0, ∞) em uma partição; quanto mais próximo de zero, mais compactos (coesos) são os grupos obtidos (Eq. (5)).

B. Índice de Xie-Beni

O Índice de Xie-Beni (Xie e Beni, 1991) é conhecido por ser a razão entre compactação e separação, isto é, ele avalia simultaneamente quão compactos e separados são os grupos em uma partição [0, ∞):

$$xb = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij} ||\mathbf{c}_j - \mathbf{x}_i||^2}{n \, . \, min_{k,j} ||\mathbf{c}_k - \mathbf{c}_j||^2} \quad (6)$$

O numerador é dado pela Soma do Erro Quadrático, e o denominador é dado pelo produto entre a quantidade de objetos da base de dados (*n*) e a menor distância entre dois protótipos. Quanto mais próximo de zero, mais compactos e separados são os grupos.

C. Índice da Silhueta Simplificada

O Índice da Silhueta (Rousseew, 1987) avalia quão compactos e separados são os grupos em uma partição, simultaneamente. Ele considera a relação de cada objeto ao grupo ao qual pertence, e também a relação do mesmo objeto ao grupo vizinho mais próximo dele. Assim, o cálculo da largura de silhueta indica seu grau de associação ao grupo ao qual pertence. Para tanto, considera-se avaliar a distância de cada objeto a todos os objetos do seu grupo, bem como a distância deste aos objetos do grupo vizinho mais próximo. Dado que esta operação é custosa computacionalmente, (Vendramin et al., 2010) propuseram substituir a dissimilaridade de cada objeto aos demais, pela dissimilaridade do objeto aos protótipos, culminando no Índice da Silhueta Simplificada:

A silhueta do objeto de índice *i* possui a forma:

$$ss(i) = \begin{cases} 1 - [a(i)/b(i)], \; se \; a(i) < b(i) \\ 0, \qquad\qquad se \; a(i) = b(i) \\ [b(i)/a(i)] - 1, se \; a(i) > b(i) \end{cases} \quad (7)$$

Em que $a(i)$ é a dissimilaridade entre o objeto $\mathbf{x}_i$ e o protótipo do grupo ao qual ele pertence, e $b(i)$ a dissimilaridade entre o objeto $\mathbf{x}_i$ e o protótipo do grupo vizinho mais próximo a ele, a dissimilaridade utilizada é a norma Euclidiana.

O índice da partição é dado pela razão entre a soma da silhueta de todos os objetos e a quantidade de objetos da base:

$$SS = \frac{1}{n} \sum_{i=1}^{n} ss(i) \quad (8)$$

O valor do índice está no intervalo [-1,1], em que quanto mais próximo de 1 melhor será a solução.

IV. Método de Ajuste de Protótipos

Nesta seção é apresentado o método que ajusta a posição de protótipos nos respectivos grupos (Szabo e Ruckl, 2021). O método objetiva identificar grupos com baixa representatividade de protótipos e, ajustá-los para aumentar a confiabilidade na análise dos grupos encontrados. O método possui três etapas:

(i) Calcular o índice de validação interna de cada grupo

(ii) Determinar a representatividade dos protótipos aos respectivos grupos a partir de (i)

Este Passo (ii) consiste em normalizar o índice obtido para cada grupo, relativamente ao melhor grupo da partição (0,1] e, em seguida, utilizar um limiar ε para determinar a necessidade de ajuste do protótipo.

A normalização do índice em cada grupo ocorre da seguinte maneira:

$$melhorIndice = arg_{max}(\mathbf{indices}) \quad (9)$$

$$\mathbf{indicesNormalizados} = \frac{\mathbf{indices}}{melhorIndice} \quad (10)$$

Para índices que objetivam a minimização do seu valor, tal como a Soma do Erro Quadrático, um passo anterior à Eq. (9) é requerido: $indices = (indices)^{-1}$.

A Eq. (10) produz índices normalizados (0,1], tal que a qualidade desejada aos grupos deva se aproximar de 1. Dessa maneira, pretende-se homogeneizar a qualidade dos grupos relativamente ao melhor grupo.

Para determinar quão representativo um protótipo é ao seu grupo, utiliza-se um limiar ε; uma constante definida no intervalo (0,1]: $if(\mathbf{indicesNormalizados}(j) < \varepsilon)$, o protótipo do grupo $\mathbf{c}_j$ deve ser ajustado.

(iii) Ajustar os protótipos, se necessário

O ajuste do protótipo é realizado inserindo um novo protótipo na proximidade daquele considerado ruim, e executando o algoritmo de agrupamento novamente. O novo protótipo é obtido, empiricamente, da seguinte maneira:

$$\mathbf{novoPrototipo} = \mathbf{\Delta prototipo} \quad (11)$$

Em que:

$$\mathbf{\Delta prototipo} = (0{,}1 * rand(1, dimensaoObjeto)) * \mathbf{prototipo} + \mathbf{prototipo}$$

Isto significa que a posição do novo protótipo inserido no conjunto é determinada por uma pequena perturbação sobre aquele a ser ajustado. O ajuste de protótipos favorece a representatividade dos mesmos e, consequentemente, a confiabilidade dos resultados.

V. Avaliação de Desempenho

A. *Metodologia*

O método de ajuste de protótipos foi avaliado para três índices de validade de grupos (Soma do Erro Quadrático, Xie-Beni e Silhueta Simplificada), em partições geradas pelo algoritmo *k*-Médias. Foram realizados três tipos de experimentos: (i) Comportamento do Método aplicado sobre uma base de dados sintética de duas dimensões (Ruspini); (ii) Análise de sensibilidade paramétrica para o Limiar ε; e (iii) Comparação entre resultado obtido pelo melhor limiar com aquele obtido antes de aplicar o Método.

O Método foi avaliado para dez bases de dados da literatura[1] e o *k*-Médias foi executado dez vezes (com inicialização de protótipos diferente em cada execução e a mesma partição avaliada para os três diferentes índices), tendo como critério de parada dez iterações do algoritmo.

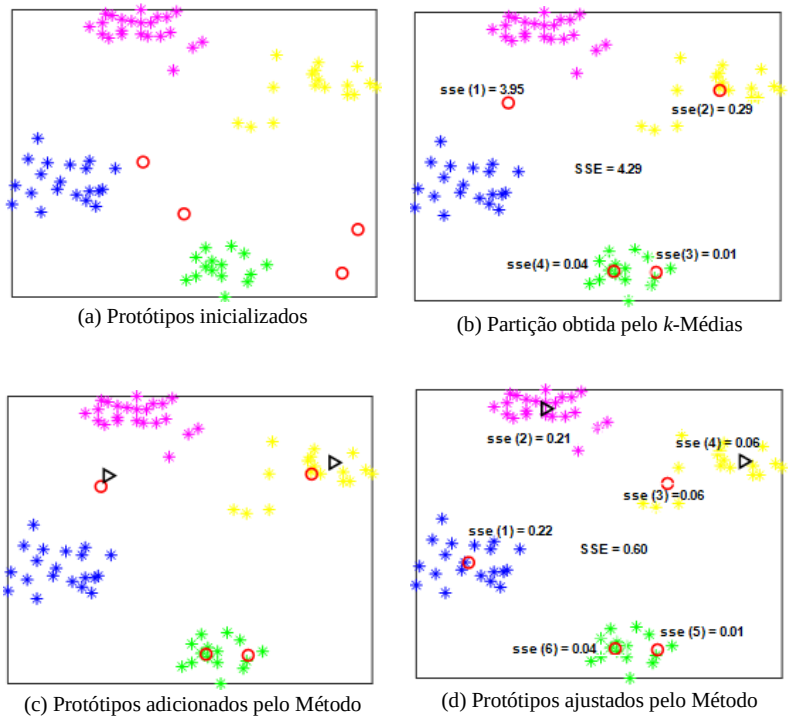
Tabela 1: Principais características das bases de dados

Bases	#Classes	#Atributos	#Objetos
Ruspini	4	2	75
Ionosphere	2	34	351
Statlog Heart	2	13	270
Haberman	2	3	306
Liver	2	7	345
Iris	3	4	150
Wine	3	13	178
Balance Scale	3	4	625
Soybean	4	35	307
Yeast	4	8	1299
Glass	6	9	214

Para a Base Yeast, apenas quatro das classes mais abundantes foram consideradas.

B. *Comportamento do Método de Ajuste de Protótipos*

O comportamento do Método foi ilustrado pela Figura 1 para a base de dados Ruspini, considerando o Índice da Soma do Erro Quadrático e o Limiar ε = 0,1.



(a) Protótipos inicializados

(b) Partição obtida pelo *k*-Médias

(c) Protótipos adicionados pelo Método

(d) Protótipos ajustados pelo Método

Figura 1: Base de dados Ruspini. Comportamento do Método em um cenário bidimensional para ε = 0,1.

[1] https://archive.ics.uci.edu/ml/datasets.php

A Fig.1(a) ilustra o cenário inicial: conjunto de dados a ser agrupado (asteriscos azuis, amarelos, roxos e verdes) e os protótipos inicializados aleatoriamente (círculos vermelhos).

A Fig.1(b) é a partição obtida pelo algoritmo *k*-Médias, tal que o Protótipo 1 representa as Classes 1 e 2 (*sse*(1)=3,95), o Protótipo 2 representa a Classe 3 (*sse*(2)=0,29), os Protótipos 3 e 4 representam a Classe 4 (*sse*(3)=0,01 e *sse*(4)=0,04). Seus valores normalizados, relativamente ao melhor grupo, estão em torno de 0.0 (*sse*(1)), 0,03 (*sse*(2)), 1,0 (*sse*(3)) e 0,25 (*sse*(4)), tal que o Protótipo 1 tem a pior representatividade e o Protótipo 3 tem a melhor representatividade.

O próximo passo é identificar protótipos com baixa representatividade, isto é, grupos para os quais o índice normalizado seja inferior ao limiar ε. Para ε=1,0, apenas o protótipo do Grupo 3 (*sse*(3)=1,0) não precisaria ser ajustado, enquanto que para ε=0,10, apenas o protótipo dos Grupos 3 e 4 não precisaria ser ajustado. Assim, para demonstrar o comportamento do Método, foi preferível ε=0,1 para evitar adição demasiada de protótipos. Portanto, o Método adiciona dois protótipos (triângulos pretos) na proximidade daqueles considerados pouco representativos, relativamente ao protótipo do melhor grupo (Fig.1(c)). Em seguida, o *k*-Médias é executado novamente para o ajuste de protótipos (Fig.1(d)). Após a aplicação do Método, verifica-se uma melhora significativa da representatividade dos grupos, medida pelo Índice da Soma do Erro Quadrático, o que reflete na qualidade da partição (*SSE*=4,29 antes da aplicação do Método e, *SSE*=0,60 após a aplicação do mesmo).

Quanto mais o valor de ε se aproxima de 1,0, mais desejado que os grupos sejam compactos, ou seja, a probabilidade de aplicar o Método aumenta e, consequentemente, maior será a quantidade de protótipos adicionados. Dessa maneira, uma análise de sensibilidade paramétrica para o limiar ε é necessária.

C. Análise de Sensibilidade Paramétrica

Foram considerados os seguintes critérios para a análise de sensibilidade paramétrica para o valor do Limiar ε: Percentual de vezes em que o método precisou ser aplicado dentre as dez execuções do algoritmo (*Aplicação do método(%)*), Eficácia do Método (*Ganho(%)*) e, Média de protótipos adicionados pelo Método (*Média de protótipos adicionados*). Para tanto, foram utilizados os seguintes valores para o limiar ε: {0,5, 0,6, 0,7, 0,8, 0,9 e 1,0}, uma vez que a possibilidade de aplicação do Método é reduzida quando ε se aproxima de 0,0.

A eficácia do Método foi verificada pelo *Ganho(%)*, o qual mede o quanto a qualidade (valor do índice de validação de agrupamento) melhorou após a aplicação do Método:

$$ganho(\%) = \left(\frac{|qualidade - qualidade\ ajustada|}{qualidade}\right) * 100\% \quad (12)$$

Os testes foram realizados para cada um dos três índices: Soma do Erro Quadrático (Fig. 2), Xie-Beni (Fig. 3) e, Silhueta Simplificada (Fig. 4), tal que a partição obtida pelo *k*-Médias em cada uma das dez execuções foi a mesma para cada índice de validação e, respectivamente, para cada valor de ε.

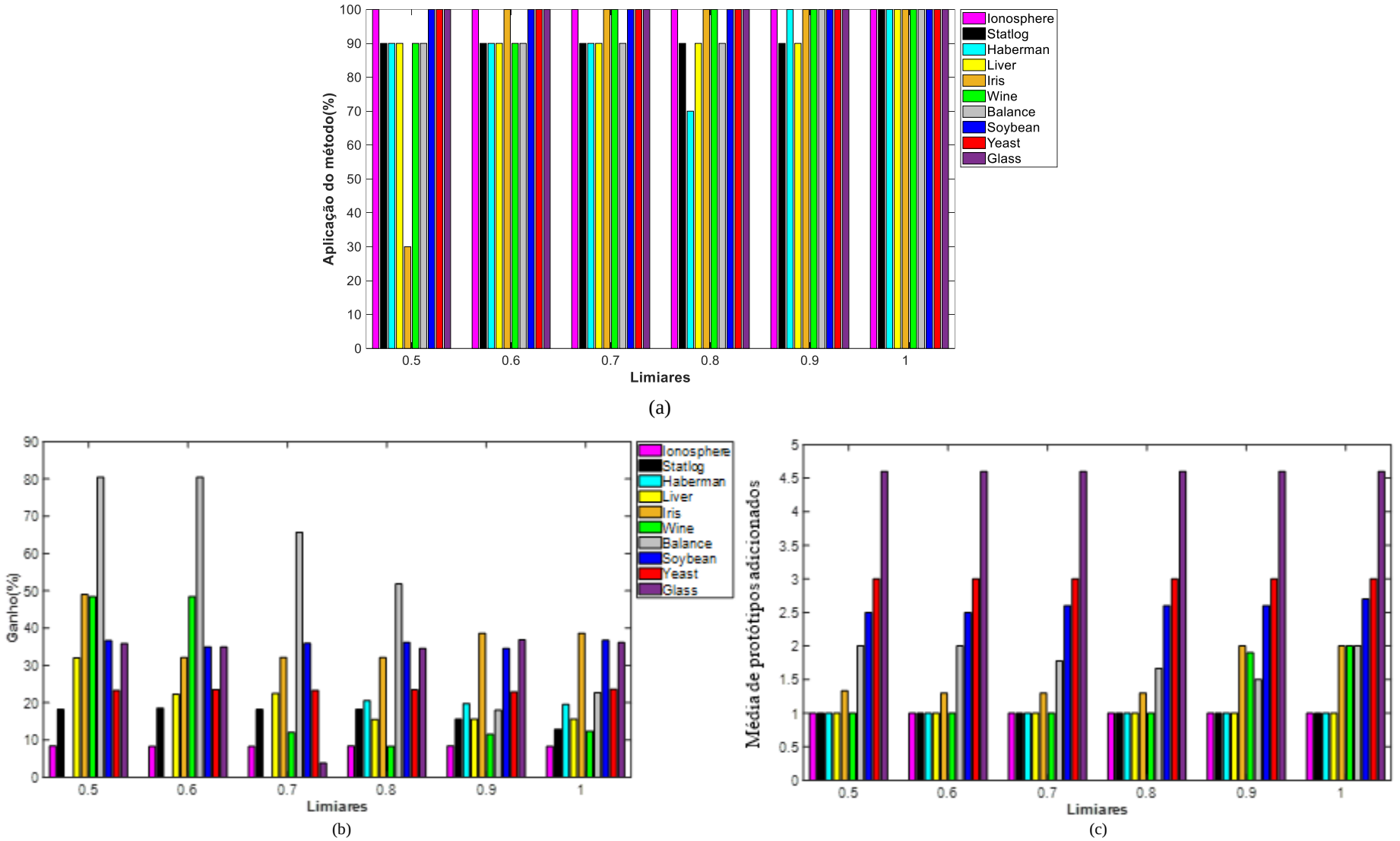


Figura 2: Soma Erro Quadrático. Percentual de vezes em que o Método foi aplicado (a), Percentual de ganho ao aplicar o Método (b) e, Média de protótipos adicionados pelo Método (c)

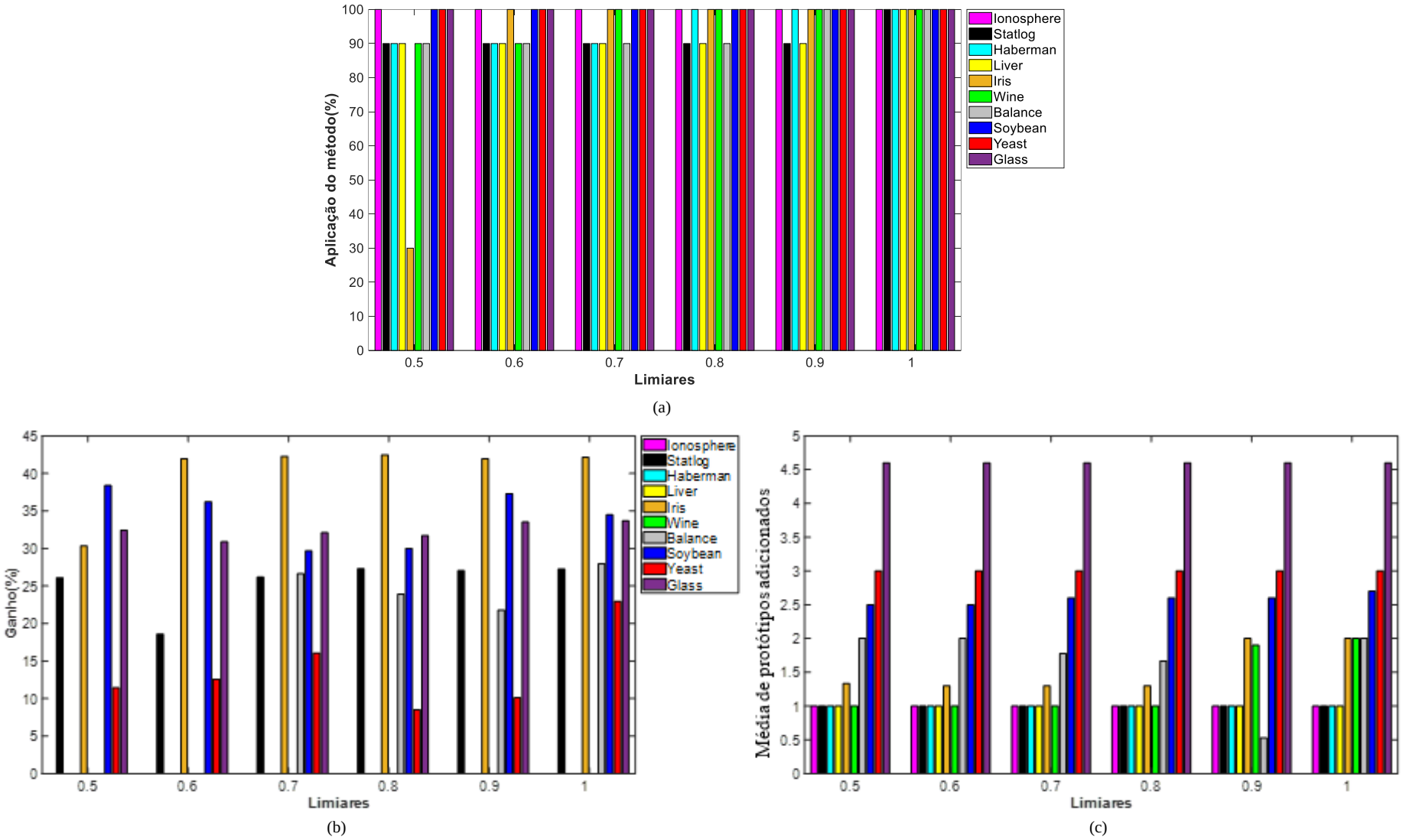
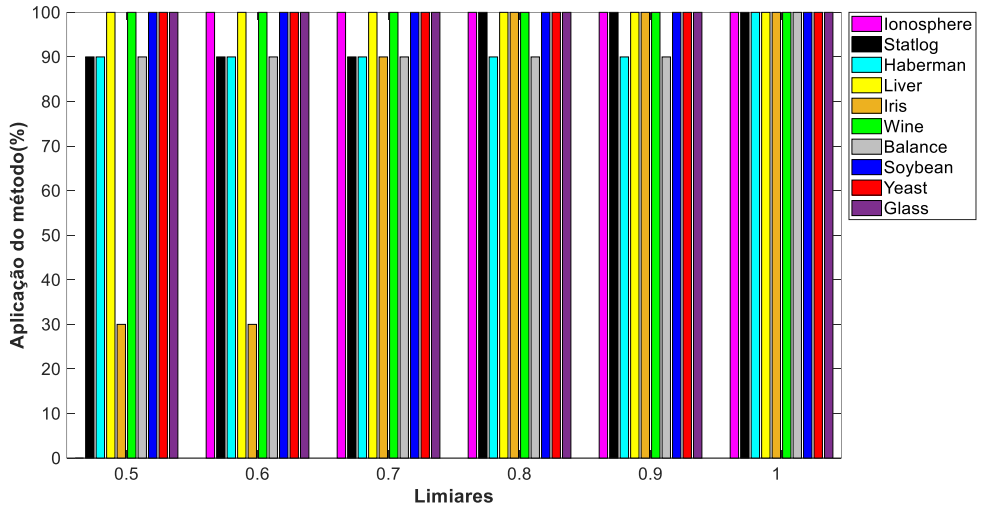
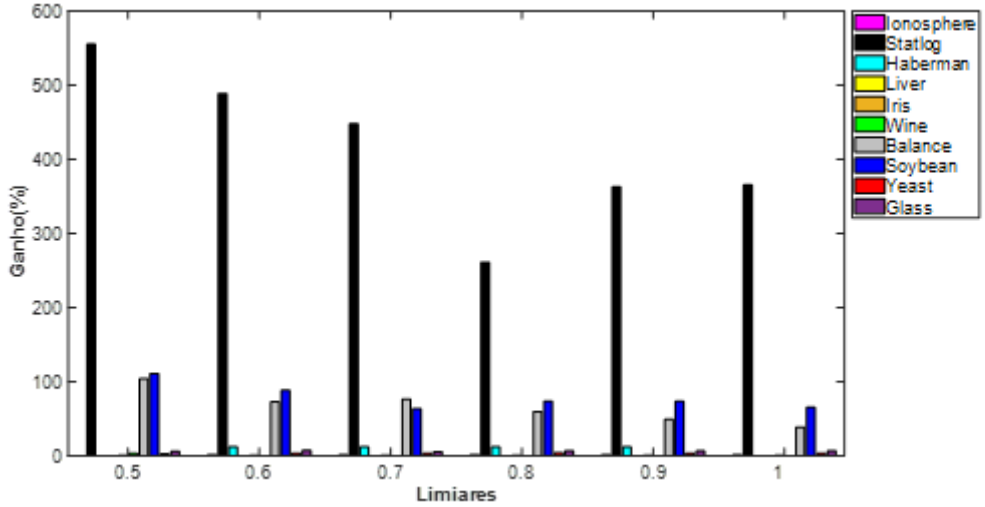


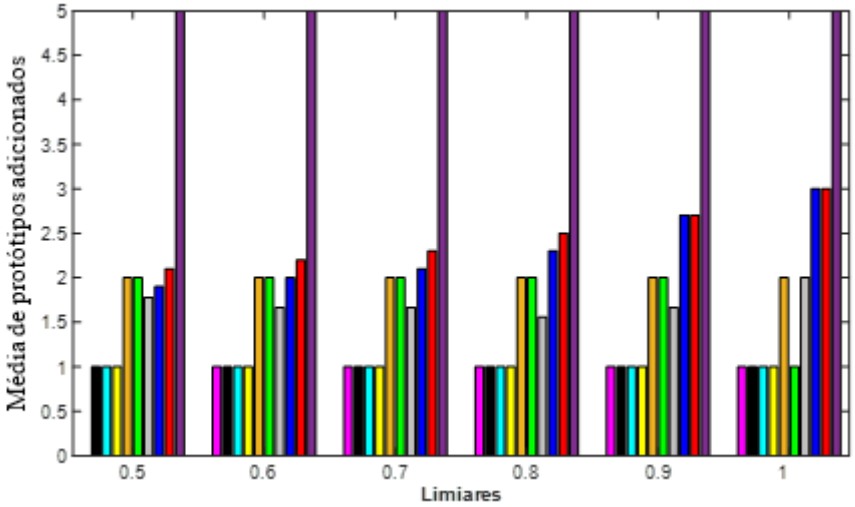
Figura 3: Xie-Beni. Percentual de vezes em que o Método foi aplicado (a), Percentual de ganho ao aplicar o Método (b) e, Média de protótipos adicionados pelo Método (c)



(a)



(b)



(c)

Figura 4: Silhueta Simplificada. Percentual de vezes em que o Método foi aplicado (a), Percentual de ganho ao aplicar o Método (b) e, Média de protótipos adicionados pelo Método (c)

Em experimentos realizados para o Índice da Soma do Erro Quadrático, observou-se que o Método foi aplicado em pelo menos 30% das execuções (Fig. 2(a)) para todos os limiares, em todas as bases de dados. Para os limiares 0,5, 0,6 e 0,7, o Método não apresentou *Ganho(%)* para ao menos uma das bases de dados (Fig. 2(b)). Entretanto, para estes três limiares, o Método produziu *Ganho(%)* significativo para as bases Liver, Wine e Balance, relativamente aos demais limiares. Para os limiares 0,8, 0,9 e 1,0 o *Ganho(%)* produzido pela aplicação do Método foi similar entre os mesmos para as respectivas bases de dados, exceto para a base de dados Balance, para a qual o limiar 0,8 foi o mais interessante, em média. Com relação à média de protótipos adicionados pelo Método (Fig. 2(c)), a variação entre os limiares não foi significativa, em média, sendo os maiores valores observados para os limiares 0,9 e 1,0, para as bases Iris e Wine. Assim, para o Índice da Soma do Erro Quadrático, o limiar mais interessante é o de 0,8, para o qual o Método produz *Ganho(%)* considerável em relação aos demais limiares, com a adição de menos protótipos, em média.

O Método não produziu resultados satisfatórios para o Índice de Xie-Beni (Fig. 3(b)) e foi aplicado em pelo menos 90% das execuções para todas as bases de dados (Fig. 3(a)); exceto para base Iris, para a qual o Método foi necessário em 30% das execuções. E a quantidade de protótipos adicionados às bases de dados se manteve entre os limiares, em média, exceto para os limiares 0,8 e 0,9, em que a quantidade de protótipos adicionados foi diferente para as bases Iris e Balance. Dentre os limiares, o de 0,7 e o de 1,0 foram os que produziram maior *Ganho(%)*. Entretanto, a quantidade de protótipos adicionados para ε=1,0 foi maior, em média. O Método não apresenta comportamento desejado para o Índice de Xie-Beni porque a adição de protótipo na vizinhança daquele a ser ajustado contribui para reduzir o denominador do índice e, consequentemente, aumentar o seu valor.

Dentre os índices investigados, o da Silhueta Simplificada foi o que apresentou o pior desempenho para o Método (Fig. 4(b)). Embora o Método tenha sido aplicado a todas as bases de dados em ao menos 30% das execuções para todos os limiares ((Fig. 4(a))), o *Ganho(%)* foi observado para até 60% das bases de dados, dentre os limiares e, diferentemente dos Índices da Soma do Erro Quadrático e de Xie-Beni, o *Ganho(%)* foi considerável para as bases de dados Statlog, Balance e Soybean, superando 100% de Ganho(%), em média. O limiar mais interessante para este índice é o 0,6. A média de protótipos foi mantida, em média, dentre os limiares para as respectivas bases de dados (Fig. 4(c)), exceto para as bases Soybean e Yeast, para as quais a quantidade de protótipos aumentou, embora não significativamente, a partir do limiar 0,6, e uma redução significativa de protótipos adicionados para a base Wine para o limiar 1,0. Assim como o Índice de Xie-Beni, o da Silhueta Simplificada também considera no seu cálculo a distância intergrupo. Portanto, índices que fazem uso de distância intergrupo não são interessantes para o Método, uma vez que a adição de protótipo na proximidade daqueles que apresentam baixa representatividade reduz a distância entre grupos.

D. Avaliação do Método para o Melhor Limiar

Esta seção compara os resultados produzidos pelo Método para o melhor dentre os limiares avaliados, para cada índice de validação de agrupamento, com os resultados produzidos pelo *k*-Médias antes da aplicação do Método na respectiva execução. Foi considerada a média dos protótipos adicionados pelo método em relação à quantidade de classes presentes nas respectivas bases de dados (*#NP*), *Ganho(%)* em relação ao índice obtido antes da aplicação do método, e o percentual de vezes em que o Método precisou ser aplicado dentre dez execuções (*Exec(%)*). Foi considerado o resultado da execução para a qual foi necessária a aplicação do Método, apenas.

A Tabela 2 a seguir apresenta o desempenho do Método para o índice da Soma do Erro Quadrático para o limiar ε=0,8.

Tabela 2: Soma do Erro Quadrático – Média de protótipos adicionados pelo Método (#NP), Ganho(%), Aplicação do Método (Exec(%)) e a Soma do Erro Quadrático (*SSE*)

		Melhor Limiar: $\varepsilon = 0{,}8$				*k*-Médias
Bases	#Classes	#NP	Ganho(%)	Exec(%)	*SSE*	*SSE*
Ionosp	2	1,00	8,37	100,00	576,28	628,90
Statlog	2	1,00	18,20	90,00	479,02	323,92
Haberman	2	1,00	20,54	70,00	97,06	25,31
Liver	2	1,00	15,50	90,00	25,35	29,00
Iris	3	1,30	32,10	100,00	5,26	8,19
Wine	3	1,00	8,30	100,00	44,91	48,98
Balance	3	1.67	51,82	90,00	105,69	219,34
Soybean	4	2.60	36,17	100,00	60,47	96,61
Yeast	4	3,00	23,45	100,00	62,84	82,33
Glass	6	4.60	34,52	100,00	15,50	23,67

Para o Índice da Soma do Erro Quadrático, o Método foi aplicado em 100% das execuções para 60% das bases de dados, isto é, ao menos um dos grupos da partição produzida pelo *k*-Médias, nas dez execuções, possuía baixa representatividade de protótipo em 60% das bases. Ao aplicar o Método, a quantidade de protótipos adicionados foi entre 33,33% (Wine) e 76,67% (Glass), e para 40% das bases o percentual foi de 50%. A qualidade da partição obtida pelo Método foi, para o limiar 0,8, superior a 8,30% para as bases avaliadas, sendo o maior valor de *Ganho(%)* observado para as bases Balance (51,82%), Soybean (36,17%), Glass (34,52%) e Iris (32,10%). Em média, o valor da Soma do Erro Quadrático (*SSE*) da partição produzida pelo Método foi melhor do que aquela antes da aplicação do mesmo para todas as bases de dados, exceto para as bases Statlog e Haberman, para as quais o Método nem sempre produz *Ganho(%)*.

A Tabela 3 apresenta o desempenho do Método para o índice da Xie-Beni para o limiar ε=0,7.

Tabela 3: Xie-Beni – Média de protótipos adicionados pelo Método (#NP), Ganho(%), Aplicação do Método (Exec(%)) e Xie-Beni (XB)

		Melhor Limiar: $\varepsilon = 0{,}7$				*k*-Médias
Bases	#Classes	#NP	Ganho(%)	Exec(%)	XB	XB
Ionosp	2	1,00	0,00	100,00	0,97	0,75
Statlog	2	1,00	26,17	90,00	0,97	1,01
Haberman	2	1,00	0,00	100,00	0,96	0,32
Liver	2	1,00	0,00	90,00	0,78	0,39
Iris	3	1,30	42,26	100,00	0,37	0,36
Wine	3	1,00	0,00	100,00	1,01	0,50
Balance	3	1,78	26,62	90,00	1,15	0,81
Soybean	4	2,60	29,68	100,00	$1{,}0*10^{17}$	$1{,}62*10^{17}$
Yeast	4	3,00	16,06	100,00	1,12	1,01
Glass	6	4,60	32,10	100,00	$0{,}37*10^{17}$	$0{,}44*10^{17}$

Para o Índice de Xie-Beni, o Método foi aplicado em 100% das execuções para 70% das bases de dados. Ao aplicar o Método, a quantidade de protótipos adicionados foi entre 43,33% (Iris) e 76,67% (Glass), e para 40% das bases o percentual foi de 50%. A qualidade da partição obtida pelo

Método foi, para o limiar 0,7, igual a zero para 40% das bases de dados, isto é, aplicando o Método, a qualidade produzida foi inferior àquela antes da aplicação do Método. Para as demais bases, o *Ganho(%)* ficou entre 16,06% (Yeast) e 42,26% (Iris). Em média, o valor de *XB* da partição produzida pelo Método foi melhor do que aquela antes da aplicação do mesmo para apenas 30% das bases de dados (Statloag, Soybean e, Glass), para as quais o valor de *XB* foi significativamente alto, tanto antes quanto depois da aplicação do Método.

A Tabela 4 a seguir apresenta o desempenho do Método para o índice da Silhueta Simplificada para o limiar ε=0,6.

Tabela 4: Silhueta Simplificada – Média de protótipos adicionados pelo Método (#NP), Ganho(%), Aplicação do Método (Exec(%)) e a Silhueta Simplificada (SS)

		Melhor Limiar: $\varepsilon = 0{,}6$				*k*-Médias
Bases	#Classes	#NP	Ganho(%)	Exec(%)	*SS*	*SS*
Ionosp	2	1,00	1,28	100,00	0,61	0,63
Statlog	2	1,00	488,59	90,00	0,56	0,14
Haberman	2	1,00	11,71	90,00	0,22	0,57
Liver	2	1,00	0,00	100,00	0,71	0,75
Iris	3	2,00	0,62	30,00	0,65	0,64
Wine	3	2,00	0,00	100,00	0,39	0,42
Balance	3	1,67	72,68	90,00	0,32	0,19
Soybean	4	2,00	87,97	100,00	0,25	0,36
Yeast	4	2,20	3,34	100,00	0,72	0,71
Glass	6	5,00	7,00	100,00	0,65	0,61

Para o Índice da Silhueta Simplificada o Método foi aplicado em 100% das execuções para 60% das bases de dados. Ao aplicar o Método, a quantidade de protótipos adicionados foi entre 50% e 83,33% (Glass). A qualidade da partição obtida pelo Método foi, para o limiar 0,6, igual a zero para 20% das bases de dados, isto é, aplicando o Método, a qualidade produzida foi inferior àquela antes da aplicação do Método. Para as demais bases o *Ganho(%)* ficou entre 0,62% (Iris) e 488,59% (Statlog). Em média, o valor de Silhueta Simplificada (*SS*) da partição produzida pelo Método foi melhor do que aquela antes da aplicação do mesmo para apenas 50% das bases de dados.

A Tabela 5 a seguir apresenta o desempenho do Método, comparativamente aos índices de validação (*SSE*(ε=0,8), *SS*(ε=0,6), *XB*(ε=0,7)), considerando o percentual de aplicação do Método em dez execuções (*Exec(%)*), o percentual de vezes em que houve ganho ao aplicar o Método (*ExecGanho(%)*), e o percentual de ganho (*Ganho(%)*).

Tabela 5: Aplicação do Método (Exec(%)), Percentual de vezes em que houve ganho (ExecGanho(%)) e Ganho(%), para SSE (ε = 0,8), XB (ε = 0,7) e SS (ε = 0,6)

	Exec(%)			ExecGanho(%)			Ganho(%)		
Bases	SSE	XB	SS	SSE	XB	SS	SSE	XB	SS
Ionosp	100,00	100,00	100,00	100,00	0,00	40,00	8,37	0,00	1,28
Statlog	90,00	90,00	90,00	33,33	33,33	100,00	18,20	26,17	488,59
Haberman	70,00	100,00	90,00	70,00	0,00	11,11	20,54	0,00	11,71
Liver	90,00	90,00	100,00	100,00	0,00	0,00	15,50	0,00	0,00
Iris	100,00	100,00	30,00	100,00	30,00	66,67	32,10	42,26	0,62
Wine	100,00	100,00	100,00	100,00	0,00	0,00	8,30	0,00	0,00
Balance	90,00	90,00	90,00	100,00	22,22	100,00	51,82	26,62	72,68
Soybean	100,00	100,00	100,00	100,00	40,00	50,00	36,17	29,68	87,97
Yeast	100,00	100,00	100,00	100,00	30,00	50,00	23,45	16,06	3,34
Glass	100,00	100,00	100,00	100,00	50,00	90,00	34,52	32,10	7,00

Para o Índice da Soma do Erro Quadrático, o Método foi aplicado em 100% das execuções para 60% das bases de dados, em 90% das execuções para 30% das bases e 70% das execuções para a base Haberman, em (*Exec(%)*). Uma vez aplicado o Método, houve algum ganho (*ExecGanho(%)*), relativamente à partição antes da sua aplicação, em 100% das vezes para 80% das bases avaliadas, produzindo *Ganho(%)* entre 8,30% (Wine) e 51,82% (Balance).

Para o Índice de Xie-Beni, o Método foi aplicado em 100% das execuções para 70% das bases de dados, e em 90% das execuções para as bases restantes (*Exec(%)*). Uma vez aplicado o Método, houve algum ganho (*ExecGanho(%)*), relativamente à partição antes da sua aplicação, em 50% das vezes para a base Glass, e nenhum ganho para 40% das bases avaliadas, produzindo *Ganho(%)* entre 16,06% (Yeast) e 42,26% (Iris).

Para o Índice da Silhueta Simplificada, o Método foi aplicado em 100% das execuções para 60% das bases de dados, sendo aplicado em 30% das execuções para a base Iris e em 90% das execuções para as bases restantes (*Exec(%)*). Uma vez aplicado o Método, houve algum ganho (*ExecGanho(%)*), relativamente à partição antes da sua aplicação, em 100% das vezes para apenas 20% das bases avaliadas, e nenhum ganho para as bases Liver e Wine. Entretanto, o Método produziu *Ganho(%)* que variou até 488,59% (Statlog).

VI. Conclusão e Trabalhos Futuros

Neste artigo foram realizados experimentos com o Método de Ajuste de Protótipos, o qual avalia a representatividade de protótipo em cada grupo, por meio de índice interno, e adiciona um protótipo na vizinhança daquele considerado ruim, ajustando-o a partir da execução do algoritmo de agrupamento. O Método foi avaliado para o algoritmo *k*-Médias, considerando experimentos individuais para três índices internos de validade de grupos: Soma do Erro Quadrático (*SSE*), Xie-Beni (*XB*) e Silhueta Simplificada (*SS*).

Foram realizados três tipos de experimentos: (i) Comportamento do Método para uma base sintética, (ii) Análise de sensibilidade paramétrica para o limiar ε e, (iii) Avaliação do Método para o melhor dentre os limiares. A partir da análise de sensibilidade (Seção V(C)), verificou-se que o valor para o melhor limiar se difere entre os índices: *SSE*(ε =0,8), *XB*(ε =0,6) e *SS*(ε =0,7), e os resultados mostraram que o *SSE* é o índice mais apropriado para o Método (Tabela 5). Dentre as dez execuções do *k*-Médias, ao menos 90% delas o Método identificou a necessidade de ajuste de protótipo para todas as bases avaliadas, produzindo *Ganho(%)* em todas as vezes em que o Método foi aplicado, para todas as bases de dados, exceto para as bases Statlog e Haberman, para as quais o percentual de vezes em que houve Ganho*(%)* ao aplicar o

Método foi de 33,33% e de 70%, respectivamente. A qualidade da partição obtida pela aplicação do Método para este índice foi entre 8,30% (Wine) e 51,82% (Balance), em média. Ao adicionar protótipos à partição, a distância intergrupo pode ser reduzida, o que pode produzir resultados pouco satisfatórios para índices que avaliam separabilidade entre grupos, tais como *XB* e *SS*.

Como propostas futuras serão investigadas a influência dos parâmetros da função que adiciona protótipo à vizinhança daquele a ser ajustado, e a determinação dinâmica do limiar ε.

REFERÊNCIAS

[1] Saiful Bahari, N. A. A., Shahidan, S., Abdullah, S. R., & Ali, N. (2017). A Preliminary Study Application Clustering System in Acoustic Emission Monitoring. MATEC Web of Conferences, 103, 02027.

[2] H. Huang, F. Meng, S. Zhou, F. Jiang, and G. Manogaran, ''Brain image segmentation based on FCM clustering algorithm and rough set,'' IEEE Access, vol. 7, pp. 12386–12396, 2019.

[3] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[4] MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.

[5] Sameh, A.S., Asoke, K.N.: Development of assessment criteria for clustering algorithms. Pattern Anal. Appl. 12(1), 79–98 (2009).

[6] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).

[7] Xie, X.L. e Beni, G. (1991). A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4), 841–846.

[8] B. Yusuf Bakhtiar, A. Bima Murti Wijaya, and H. Dwi Cahyono, "PENGEMBANGAN SISTEM ANALISIS AKADEMIS MENGGUNAKAN OLAP DAN DATA CLUSTERING STUDI KASUS : AKADEMIK UNIVERSITAS SEBELAS MARET SURAKARTA," J. Teknol. Inf. ITSmart, vol. 4, no. 1, p. 01, Sep. 2016.

[9] Vendramin, L., Campello, R.J., Hruschka, E.R.: Relative clustering validity criteria: a comparative overview. Stat. Anal. Data Min. 3(4), 209–235 (2010).

[10] Ruspini, E. H., Bezdek, J. C. & Keller, J. M. (2019). Fuzzy Clustering: A Historical Perspective. IEEE Computational Intelligence Magazine, 14(1), 45-55.

[11] J. W. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[12] Hruschka, E. R., e Ebecken, N. F. F. (2003). A genetic algorithm for cluster analysis. Intelligent Data Analysis, 7(1), 15–25.

[13] A. Szabo e F. O. de França, "The proposal of dynamic thresholds in an immune algorithm for fuzzy clustering," 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2015, pp. 1-8, doi: 10.1109/FUZZ-IEEE.2015.7338021.

[14] De Castro, L. N., e Timmis, J. (n.d.). An artificial immune network for multimodal function optimization. Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600). doi:10.1109/cec.2002.1007011.

[15] Masood, Muhammad Ali, e M. N. A. Khan. "Clustering techniques in bioinformatics." IJ Modern Education and Computer Science 1 (2015): 38-46.

[16] Popkova, E. G., Tyurina, Y. G., Sozinova, A. A., Bychkova, L. V., Zemskova, O. M., Serebryakova, M. F., & Lazareva, N. V. (2017). Clustering as a growth point of modern Russian business. In Integration and Clustering for Sustainable Economic Growth (pp. 55-63). Springer, Cham.

[17] Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.

[18] Szabo, A., e Ruckl, T. (2021). Improving Prototypes Representatividade by Internal Validity Index Analysis. Learning and Nonlinear Models. Aceito para publicação.