Universidade Federal de Pernambuco
Centro de Informática

Graduação em Engenharia da Computação

Block Clustering K-means para Variáveis Simbólicas do Tipo Intervalo

Gibson Belarmino Nunes Barbosa

Trabalho de Graduação

Recife
19 de setembro de 2013

Universidade Federal de Pernambuco
Centro de Informática

Gibson Belarmino Nunes Barbosa

Block Clustering K-means para Variáveis Simbólicas do Tipo Intervalo

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientador: *Prof. Francisco de Assis Tenório de Carvalho*

Recife
19 de setembro de 2013

Resumo

Em análise de agrupamentos o objetivo é agrupar os dados de entrada, de forma que indivíduos no mesmo grupo devem ter similaridade maior que indivíduos em grupos diferentes. Um dos algoritmos mais conhecidos com esse propósito é o K-means.

Enquanto no algoritmo de partição K-means tradicional temos como resultado uma partição de indivíduos, o algoritmo de Block Clustering K-means traz simultaneamente uma partição de indivíduos e outra de variáveis. Desse modo, neste trabalho é feito um estudo do modelo de Block Clustering K-means para dados quantitativos e é proposto e analisado o algoritmo de Block Clustering K-means para variáveis simbólicas do tipo intervalo.

Palavras-chave: Block Clustering, K-means, dados quantitativos, dados de tipo Intervalo, Análise de agrupamentos

Sumário

CAPÍTULO 1

Introdução

Análise de agrupamento é um método muito utilizado em mineração de dados e reconhecimento de padrões, que visa distribuir dados em grupos, subgrupos ou categorias, de acordo com o princípio que indivíduos no mesmo grupo tenham grau de similaridade mais elevado que indivíduos em grupos diferentes. As técnicas de agrupamento mais utilizadas são os métodos hierárquicos e os métodos particionais [8]. Os métodos hierárquicos geram grupos de forma que a partição vai sendo gerada intermediariamente e novas partições geradas são influenciadas por partições anteriores, desenvolvendo-se uma hierarquia de partições. Já os métodos particionais, o qual é utilizado nesse trabalho, geram uma partição dos dados de entrada em um número fixo, definido previamente, de grupos que são, em geral, gerados a partir da otimização de uma função objetivo. Para esse trabalho as medidas de dissimilaridade são feitas através da distância Euclidiana, dada sua grande utilização na maioria dos algoritmos de agrupamento particionais e que se comporta bem para bases de dados que trazem uma distribuição aproximadamente hiperesférica e são linearmente separáveis.

Para uma descrição mais precisa do mundo real através de informações que possam ser processadas computacionalmente deve-se levar em conta a variabilidade e/ou incerteza dos dados, para tanto os dados podem ser descritos por conjuntos de categorias, histogramas, intervalos, distribuições de pesos, etc [3]. Para este trabalho utilizamos dados do tipo intervalo, que levam em consideração o intervalo de possíveis valores para determinada variável, assim, poderíamos, por exemplo, descrever melhor o clima em uma cidade qualquer que variaria de 15°C a 29°C.

Enquanto alguns métodos têm como objetivo gerar partições de indivíduos com o máximo de precisão e outros objetivam gerar partições de variáveis, métodos de agrupamento baseados em blocos (ou block clustering methods) consideram a formação do grupo de indivíduos e de variáveis simultaneamente, organizando os dados em blocos homogêneos. Como exemplo, na figura 1.1 pode ser vista uma matriz 8x8 onde as linhas podem ser representadas como os indivíduos e as colunas como as variáveis, assim seriam feitas permutações entre as linhas e entre as colunas de forma que tanto as linhas como as colunas com valores mais similares ficassem mais próximos (no mesmo grupo), formando, assim, blocos que combinam cada grupo de indivíduos com cada grupo de variáveis. Com essas partições é possível gerar conjuntos de dados menores que o inicial, para serem processados com menos tempo computacional. Modos de utilização do método de agrupamento simultâneo podem ser vistos em [5], [6], [7].

	a	b	c	d	e	f	g	h
A	1	2	3	1	2	2	1	3
B	1	2	3	1	2	2	1	3
C	4	5	6	4	5	5	4	6
D	4	5	6	4	5	5	4	6
E	1	2	3	1	2	2	1	3
F	4	5	6	4	5	5	4	6
G	1	2	3	1	2	2	1	3
H	4	5	6	4	5	5	4	6

(a)

	g	d	a	b	e	f	c	h
G	1	1	1	2	2	2	3	3
A	1	1	1	2	2	2	3	3
B	1	1	1	2	2	2	3	3
E	1	1	1	2	2	2	3	3
D	4	4	4	5	5	5	6	6
H	4	4	4	5	5	5	6	6
F	4	4	4	5	5	5	6	6
C	4	4	4	5	5	5	6	6

(b)

Figura 1.1 Matriz representativa de uma base de dados onde as linhas são os indivíduos e as colunas são as variáveis, em (a) temos os dados iniciais desorganizados e em (b) os dados após a formação dos blocos por similaridade dos valores

Esse trabalho propõe a técnica de Block Clustering K-means (BCKM) para dados simbólicos no contexto de dados do tipo intervalo com medidas de dissimilaridade entre centroides e padrões utilizando distância Euclidiana, assim tal técnica tem como objetivo fazer o particionamento simultâneo de indivíduos e variáveis. Com isso, o modelo em estudo traz como vantagem, quando comparado ao modelo K-means (KM) convencional (sem blocos), a informação adicional do particionamento das variáveis, sem que haja degradação do desempenho na maioria dos casos. Mostraremos, também, que a escolha da configuração de grupos de variáveis traz influência no desempenho, consequentemente uma melhor escolha da quantidade de grupos de variáveis para uma quantidade fixa de grupos de indivíduos pode melhorar, ou no mínimo, igualar a performance obtida pelo modelo K-means tradicional.

Estudos vêm sendo desenvolvidos visando o agrupamento de dados simbólicos com o mínimo de custo computacional e taxas de acertos mais altas [1], [9]. Outros estudos propõem técnicas de agrupamento com blocos utilizando simultaneamente grupos de indivíduos e variáveis [5], [6].

Buscamos com esse trabalho fazer um estudo da técnica de Block Clustering para dados quantitativos utilizando distância euclidiana, e assim propor a técnica de block clustering para dados simbólicos no contexto de dados do tipo intervalo utilizando distância Euclidiana, que não foi explorada, fazendo uma análise do seu desempenho através do índice da taxa de erro de classificação global OERC (Overall Error Rate of Classification) e do índice de Rand Corrigido CR (Corrected Rand Index), e comparando com o algoritmo de agrupamento de dados k-means tradicional, sem a utilização de blocos. Mostramos, também, a sua utilização para classes com padrões tendo diferentes níveis de homogeneidade nas variáveis.

Este trabalho está organizado da seguinte maneira. O capítulo 2 traz uma introdução ao algoritmo de partição k-means e sua aplicação para variáveis de tipo intervalo. No capítulo 3 é explorada a abordagem de Block Clustering para dados quantitavos, sendo mostrado o seu algoritmo e sua aplicação em diversas bases de dados, sendo algumas simuladas e outras reais.

O capítulo 4 é focado no algoritmo de Block Clustering para dados de tipo intervalo, assim seu algoritmo é apresentado e depois há a análise experimental para bases de dados de tipo real mostrando a eficiência do método proposto. Por fim, o trabalho é concluído no capítulo 5.

CAPÍTULO 2

O algoritmo de partição K-means

Proposto há mais de 50 anos [8], o K-means é o algoritmo de partição mais utilizado em agrupamento de dados dada sua simplicidade de implementação e eficiência.

O K-means tem esse nome devido a divisão dos indivíduos em K grupos onde cada grupo é representado por um protótipo que é a média do somatório dos valores de cada padrão inserido no grupo, então temos K protótipos.

A distância entre cada protótipo e os indivíduos no grupo pode ser calculada através de diversas métricas como a diferença quadrática (distância euclidiana), a qual é utilizada neste trabalho, ou o módulo das diferenças(distância city-block). Assim o algoritmo de partição K-means tem como objetivo encontrar partições onde o somatório destas distâncias seja menor possível.

O número K de grupos deve ser definido à priori e sua escolha implica num melhor ajuste dos indivíduos, também deve ser definida a partição inicial que, idealmente, deve ser feita para várias inicializações do algoritmo, já que as partições finais obtidas dependem fortemente da configuração inicial definida, então com varias inicializações é aumentada a probabilidade de atingirmos o agrupamento ideal para os indivíduos. A seguir podemos ver a descrição de alto nível do algoritmo K-means:

1) Selecione uma partição inicial com K clusters;

2) Definir os protótipos calculando as médias dos valores representativos dos indivíduos em cada grupo;

3) Calcular a distância de todos os indivíduos para cada um dos K protótipos e inserir cada indivíduo no grupo ao qual ele obteve a menor distância;

4) Se nenhum indivíduo mudou de grupo termine, caso contrário repita a partir do item (2)

2.1 K-means para dados do tipo intervalo

Existem situações práticas em que a descrição dos indivíduos a partir de um único valor para cada uma de suas variáveis não é suficiente. Um exemplo é o caso da temperatura das cidades, que não são descritas por exatamente um valor, mas sim por um intervalo possível de valores onde a temperatura pode variar, uma cidade qualquer pode ter esse intervalo entre 6°C e 18°C por exemplo, assim particularmente falando às vezes as variáveis de um indivíduo necessitam de mais de um valor para serem descritas e esse tipo de dado é tipicamente chamado de 'dado simbólico' [2].

Nesse trabalho trataremos de tais dados simbólicos do tipo intervalo, assim temos o conjunto de n indivíduos $E = \{e_1, \ldots, e_n\}$ e o conjunto de p variáveis $X = \{x_1, \ldots, x_p\}$, então temos a descrição de um indivíduo i como $e_i = (y_{i1}, \ldots, y_{ip})$ e $y_{ij} = [a_{ij}, b_{ij}](i = 1, \ldots, n; j = 1, \ldots, p)$ é o intervalo de valores que cada variável de um indivíduo pode assumir, tendo a_{ij} como o valor mínimo e b_{ij} como o valor máximo. Podemos perceber que para $a_{ij} = b_{ij}$ temos uma variável de valor único, sendo, então, do tipo real.

Como se trata de algoritmo de partição K-means, proposto para variáveis do tipo intervalo em [3], temos um número K de grupos em uma partição $P = (P_1, \ldots, P_K)$ para o conjunto $\Omega = 1, \ldots, n$, e consideramos uma matriz U onde $U_{ik}(i = 1, \ldots, n)(k = 1, \ldots, K)$ indica a pertinência do indivíduo e_i ao grupo P_k, assim $U_{ik} = 1$ para $x_i \in P_k$ e $U_{ik'} = 0$ para qualquer outro grupo $P_{k'}(k' = 1, \ldots, K)$ e $k' \neq k$. Os representantes de cada grupo, no caso os protótipos, são descritos pelo sistema $G = (g_1, \ldots, g_K)$ e $g_k = (g_{k1}, \ldots, g_{kp})$ para $g_{kj} = [\alpha_{kj}, \beta_{kj}](j = 1, \ldots, p)(k = 1, \ldots, K)$ que devem ser calculados de forma a otimizar (minimizar) o critério de adequação J localmente.

Descrição do índice de adequação J, para variáveis de tipo intervalo:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{p} (U_{ik})[(a_{ij} - \alpha_{kj})^2 + (b_{ij} - \beta_{kj})^2] \tag{2.1}$$

Os protótipos são calculados da seguinte forma:

$$\alpha_{kj} = \frac{\sum_{i=1}^{n} a_{ij}}{n} \quad e \quad \beta_{kj} = \frac{\sum_{i=1}^{n} b_{ij}}{n} \tag{2.2}$$

1) *Inicialização:*
   Aleatoriamente construir a partição inicial $P^{(0)} = (P_1^{(0)}, \ldots, P_K^{(0)})$ de indivíduos, onde U_{ik} deve corresponder a essa partição.

2) *Calcular protótipos*
   Computar os protótipos $g_{kj} = [\alpha_{kj}, \beta_{kj}](j = 1, \ldots, p)(k = 1, \ldots, K)$ de acordo com a equação (2.2):
   $$\alpha_{kj} = \frac{\sum_{i=1}^{n} a_{ij}}{n} \quad e \quad \beta_{kj} = \frac{\sum_{i=1}^{n} b_{ij}}{n}$$

3) *Definir melhor partição*

   teste = 0

   Para $i = 1, \ldots, n$ fazer

   $k_i = k$, tal que $U_{ik} = 1$

   $$k_f = \underset{1 \leq k \leq K}{\operatorname{argmin}} \sum_{j=1}^{p} [(a_{ij} - \alpha_{kj})^2 + (b_{ij} - \beta_{kj})^2]$$

se $k_i \neq k_f$, então

$teste = 1$

$U_{ik_f} = 1$

$U_{ik_i} = 0$

4) *Critério de parada*
If teste = 0 então PARE, caso contrário volte ao item (2)

CAPÍTULO 3

Block Clustering K-means para Dados Quantitativos

Nessa seção fazemos análises experimentais para diversas bases de dados sintéticas e reais aplicadas ao algoritmo de Block Clustering K-means(BCKM) para dados quantitativos, que foi apresentado em [10].

O algoritmo traz como objetivo particionar um conjunto de padrões em K grupos $P_1,\ldots,P_K$ e particionar as variáveis correspondentes aos padrões em Q grupos $H_1,\ldots,H_Q$, fornecendo, também, os protótipos $g_{kq}(k=1,\ldots,K)(q=1,\ldots,Q)$ que são calculados de forma a otimizar (minimizar) o critério de adequação J, assim teremos um melhor ajuste entre os grupos e os protótipos. Abaixo é descrito o critério J:

$$J=\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})(y_{ij}-g_{kq})^2 \tag{3.1}$$

onde $U_{ik}(i=0,\ldots,n)(k=0,\ldots,K)$ é uma posição da matriz U que indica a pertinência do indivíduo e_i ao grupo P_k, assim $U_{ik}=1$ para $e_i\in P_k$ e $U_{ik'}=0$ para qualquer outro grupo $P_{k'}(k'=1,\ldots,K)$ e $k'\neq k$, já na matriz V temos $V_{jq}(j=0,\ldots,p)(q=0,\ldots,Q)$ indicando a pertinência da variável x_j ao grupo de variável H_q, então $V_{jq}=1$ para $x_j\in H_q$ e $V_{jq'}=0$ para qualquer outro grupo $H_{q'}(q'=1,\ldots,Q)$ e $q'\neq q$. E $e_i=(y_{i1},\ldots,y_{ip})(i=1,\ldots,n)(j=1,\ldots,p)$ é a descrição do indivíduo.

Os protótipos são computados da seguinte maneira:

$$g_{kq}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})} \tag{3.2}$$

Os indivíduos devem ser alocados nos grupos de indivíduos de forma que o critério J seja minimizado, assim o indivíduo e_i deve ser alocado ao grupo k seguindo o seguinte critério:

$$k=\operatorname*{argmin}_{1\leq l\leq K}\sum_{q=1}^{Q}\sum_{j=1}^{p}(V_{jq})(y_{ij}-g_{lq})^2 \tag{3.3}$$

As variáveis devem ser alocados em grupos de variáveis de forma que o critério J seja minimizado, assim a variável x_j deve ser alocado ao grupo q seguindo o seguinte critério:

$$q = \underset{1\leq l\leq Q}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{n}(U_{ik})(y_{ij}-g_{kl})^2 \tag{3.4}$$

3.1 Descrição do algoritmo

Os seguintes passos descrevem o algoritmo Block Clustering K-Means(BCKM) para dados quantitativos de forma que possa ser implementado em qualquer linguagem de programação:

1) *Inicialização:*

   Definir o número K de grupos de indivíduos e o número H de grupos de variáveis.

   Aleatoriamente construir a partição inicial $P^{(0)} = (P_1^{(0)},\ldots,P_k^{(0)},\ldots,P_K^{(0)})$ de indivíduos, U_{ik} deve estar de acordo com essa partição, e a partição inicial $H^{(0)} = (H_1^{(0)},\ldots,H_q^{(0)},\ldots,H_Q^{(0)})$ de variáveis e V_{jq} deve corresponder a essa partição.

2) *Passo 1: Calcular protótipos*

   Computar os protótipos $g_{kq}(k = 1,\ldots,K)(q = 1,\ldots,Q)$ de acordo com a equação (3.2):

$$g_{kq} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})}$$

3) *Passo 2: Definir melhor partição de indivíduos*

   teste = 0
   Para $i = 1,\ldots,n$ fazer

   $k_i = k$ onde $U_{ik} = 1$

   k_f é obtido de acordo com a equação (3.3), assim:

$$k_f = \underset{1\leq l\leq K}{\operatorname{argmin}} \sum_{q=1}^{Q}\sum_{j=1}^{p}(V_{jq})(y_{ij}-g_{lq})^2$$

   se $k_i \neq k_f$, então

   teste = 1

   $U_{ik_f} = 1$

   $U_{ik_i} = 0$

4) *Passo 3: Calcular protótipos*

   Computar os protótipos $g_{kq}(k = 1,\ldots,K)(q = 1,\ldots,Q)$ de acordo com a equação (3.2) para a nova partição U_{ik} calculada no passo anterior, assim:

$$g_{kq} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})}$$

5) *Passo 4: Definir melhor partição de variáveis*

   teste2 = 0

   Para $j = 1,\ldots,p$ fazer

   $q_i = q$ onde $V_{jq} = 1$

   q_f é obtido de acordo com a equação (3.4), assim:

$$q_f = \underset{1\leq l\leq Q}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{n}(U_{ik})(y_{ij} - g_{kl})^2$$

   se $q_i \neq q_f$, então

   teste2 = 1

   $V_{jq_f} = 1$

   $V_{jq_i} = 0$

6) *Critério de parada*

   Se teste = 0 e teste2 = 0 então PARE, caso contrário volte ao item (2)

3.2 Análise experimental

Com a finalidade de analisar o comportamento do algoritmo de block clustering fazemos diversos experimentos com bases de dados criadas sinteticamente e com bases de dados reais. Para avaliar sua performance em todos os casos de teste a taxa de erro de classificação global - OERC (Overall Error Rate of Classification) - e do índice de Rand corrigido - CR (Corrected Rand Index) - são avaliados e comparados com os valores dos mesmo índices obtidos para o modelo K-means tradicional (Sem blocos). Para todas as configurações de teste, seja nas bases de dados sintética ou real, o algoritmo de block clustering para dados quantitativos foi aplicado 1000 vezes e foi escolhida a repetição que obteve o menor critério de adequação J (equação 3.1).

O índice OERC indica o quanto a partição obtida ao fim da execução do algoritmo está de acordo com partição a priori, sendo um valor entre 0 e 1 onde quanto mais próximo de 0 for o índice OERC, melhor é a partição obtida.

O índice CR fornece o grau de similaridade entre a partição obtida pela execução do algoritmo e a partição a priori, não sendo sensível a quantidade de classes ou a distribuição de items

nos clusters. Ele assume valores entre -1 e 1, onde quanto mais próximo de 1 indica melhor correspondencia entre as partições, já valores menores indicam maior aleatoriedade.

3.2.1 Conjunto de dados sintéticos

São feitos experimentos para 4 configurações diferentes de bases de dados sintéticas, que são geradas através de uma distribuição Gaussiana bivariada, de acordo com [4], com os parâmetros específicos a cada configuração. Tais bases sitéticas são geradas controladamente de forma a ser analisado o comportamento do modelo BCKM em determinadas situações. A primeira e a segunda configuração possuem duas variáveis e duas classes, se diferenciando na homogeneidade, de forma que na primeira uma das classes é mais homogênea em um das variáveis e a outra classe é mais homogênea na outra variável (figura 3.1), enquanto para a segunda configuração as duas classes são mais homogêneas em uma das variáveis e heterogêneas na outra variável (figura 3.2), já para a terceira configuração consideramos 3 variáveis e duas classes que são claramente separadas por uma das variáveis enquanto as outras duas sobrepõem as classes (figura 3.3), e na quarta configuração temos 4 variáveis e duas classes que em duas das variáveis são bem separadas e nas outras duas são sobrepostas (figura 3.4). A quantidade de clusters de indivíduos a ser formado é de acordo com o número de classes à priori, no caso serão dois clusters, e é pedido dois clusters de variáveis para todas as configurações testadas.

3.2.1.1 Conjunto de dados sintéticos 1

A primeira configuração de dados sintéticos possui 200 padrões descritos por duas variáveis e que se dividem em duas classes com 100 indivíduos cada, eles foram gerados através dos parâmetros descritos na tabela 3.1 de forma que a classe 1 seja mais homogênea para a variável x_1 e pouco homogênea pra variável x_2, enquanto a classe 2 tenha homogeneidade maior para variável x_2 e menor para variável x_1 e tendo uma intercepção entre as classes. A configuração dos dados pode ser vista pela figura 3.1.

Tabela 3.1 Parâmetros de configuração do conjunto de dados sintéticos 1

	Classe 1		Classe 2	
	Média	Variância	Média	Variância
x_1	0	0.1	0	2
x_2	2	2	0	0.1

Tabela 3.2 Performance dos algoritmos aplicados ao conjunto de dados sintéticos 1

	OERC	CR
BCKM	0.135	0.530
KM	0.135	0.530

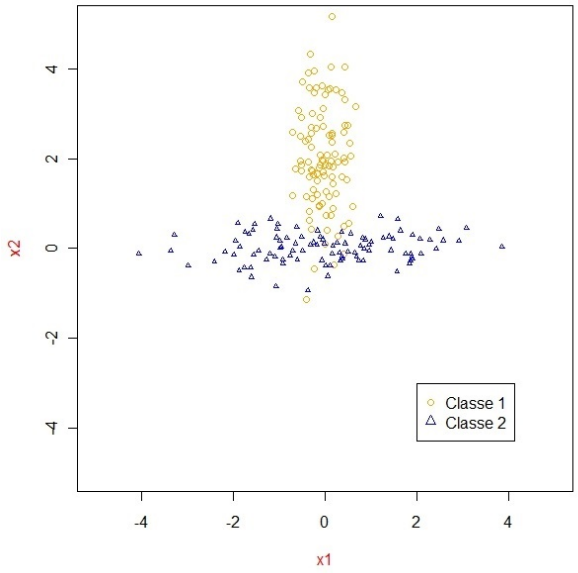

Figura 3.1 Conjunto de dados Sintético 1

Tabela 3.3 Matriz de confusão obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 1

Classe	P_1	P_2	Somatório
1	27	73	100
2	100	0	100
Somatório	127	73	200

Tabela 3.4 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 1

	H_1	H_2
P_1	**27.9591**	237.2700
P_2	58.2185	**8.5217**

Tabela 3.5 Protótipos obtidos para cada um dos blocos para o conjunto de dados sintéticos 1

	H_1	H_2
P_1	0.1481	-0.0099
P_2	2.5894	-0.0321

Após aplicarmos os algoritmos de agrupamento BCKM e KM para dados quantitativos na configuração de dados descrita, obtivemos os índices OERC e CR de acordo com a tabela 3.2, e é percebido que ambos algoritmos trazem o mesmo resultado. A variável x_1 ficou alocada em H_2 enquanto a variável x_2 foi inserida em H_1, então é possível ver através da matriz do critério de adequação (tabela 3.4) obtida, que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis, que o H_1 possui maior homogeneidade para P_1, o qual possui todos indivíduos da classe 2 (tabela 3.3) e que H_2 é mais homogêneo em P_2 que contém maior parte dos indivíduos da classe 1 (tabela 3.3). Então conseguimos distinguir a maior homogeneidade da classe 1 na variável x1 e a maior homogeneidade da classe 2 na variável x2. Pode ser visto através da matriz de blocos obtida, descrita na tabela 3.6, para os padrões de entrada a similaridade entre os padrões nos mesmo bloco e a dissimilaridade entre padrões em blocos diferentes.

Tabela 3.6 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados sintéticos 1

		H_1	H_2
		x_1	x_2
	e_3	1.3045	-0.2479
	e_{17}	0.9236	0.5929
	e_{21}	-0.1277	0.4116
	e_{25}	-1.1413	-0.4065
P_1	e_{26}	1.2597	0.0163
	e_{28}	0.7308	0.0141
	e_{30}	1.1456	-0.4055
	e_{31}	1.1877	-0.1665
	e_{35}	-0.4458	-0.2283
	e_{37}	0.4049	0.0378
	e_{42}	1.2232	0.2459
	e_1	1.9620	-0.0566
	e_2	2.2610	-0.2935
	e_4	1.7187	-0.5220
	e_5	1.8027	-0.1290
P_2	e_6	3.5824	-0.3464
	e_7	1.9446	-0.5351
	e_8	5.5693	0.7956
	e_9	4.0376	0.4413
	e_{10}	3.5237	0.0569
	e_{11}	1.7512	-0.3046

3.2.1.2 Conjunto de dados sintéticos 2

Consideramos duas variáveis e duas classes que possuem 100 indivíduos cada. Este segundo conjunto de dados sintéticos foi gerado através dos parâmetros na tabela 3.7 de forma que classe 1 e a classe 2 tivessem uma intercepção e ambas possuíssem maior homogeneidade na variável x_1, quando comparada a variável x_2, a qual os dados devem estar distribuídos mais heterogeneamente, sua distribuição pode ser vista na figura 3.2.

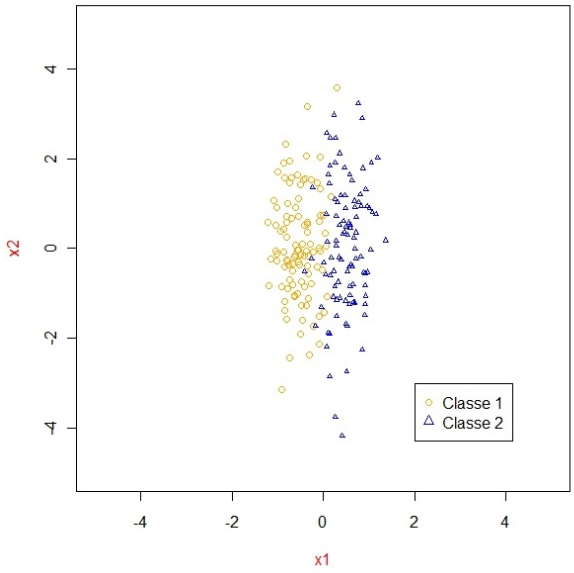

Figura 3.2 Conjunto de dados Sintético 2

Tabela 3.7 Parâmetros de configuração do conjunto de dados sintéticos 2

	Classe 1		Classe 2	
	Média	Variância	Média	Variância
x_1	-0.5	0.1	0.5	0.1
x_2	0	2	0	2

Tabela 3.8 Performance dos algoritmos aplicados ao conjunto de dados sintéticos 2

	OERC	CR
BCKM	0.460	0.001
KM	0.460	0.001

Como resposta a aplicação do algoritmo de BCKM a variável x_1 fica inserida em H_2 e a variável x_2 vai para H_1, então vemos na tabela 3.8 que os algoritmos KM e BCKM têm a mesma

performance. Pela tabela 3.11, que mostra a matriz do critério de adequação, H_2 possui mais homogeneidade para os dois clusters de indivíduos, enquanto em H_1 os cluster de indivíduos têm heterogeneidade maior. Podemos visualizar os dados de acordo com os blocos formados pela tabela 3.10 que mostra os 10 primeiros indivíduos alocados de cada um dos dois grupos de indivíduos, e perceber que indivíduos no mesmo bloco possuem valores mais semelhantes que indivíduos em blocos diferentes e a variação entre os valores dos indivíduos num mesmo bloco é menor de acordo com o critério de adequação (tabela 3.11), que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e é sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis, para aquele bloco (J_{kh}) que para indivíduos em blocos diferentes.

Tabela 3.9 Matriz de confusão obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 2

Classe	P_1	P_2	Somatório
1	59	41	100
2	51	49	100
Somatório	127	73	200

Tabela 3.10 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados sintéticos 2

		H_1	H_2
		x_2	x_1
P_1	e_1	-0.0380	-0.5566
	e_3	-0.6955	-0.7479
	e_4	-0.2813	-1.0220
	e_5	-0.1973	-0.6290
	e_7	-0.0554	-1.0351
	e_{11}	-0.2488	-0.8046
	e_{13}	-0.4684	-0.0329
	e_{14}	-0.1758	-0.3348
	e_{17}	-1.0764	0.0929
	e_{18}	-0.4095	-0.1334
P_2	e_2	0.2610	-0.7935
	e_6	1.5824	-0.8464
	e_8	3.5693	0.2956
	e_9	2.0376	-0.0588
	e_{10}	1.5237	-0.4431
	e_{12}	0.5949	-0.1002
	e_{15}	0.5837	-1.2108
	e_{16}	0.5298	-0.3515
	e_{19}	0.7430	-0.0767
	e_{20}	1.4644	-0.1427

Tabela 3.11 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 2

	H_1	H_2
P_1	75.7510	**37.2821**
P_2	52.0370	**37.3825**

Tabela 3.12 Protótipos obtidos para cada um dos blocos para o conjunto de dados sintéticos 2

	H_1	H_2
P_1	-0.8759	-0.0800
P_2	1.2140	0.0633

3.2.1.3 Conjunto de dados sintéticos 3

Para esta simulação consideramos três variáveis e 200 padrões divididos igualmente em duas classes, que são dispostas de forma que as classes sejam sobrepostas para as variáveis x_1 e x_2,

mas sejam disjuntas na variável x_3, assim podemos ver através da figura 3.3 os dados gerados com os parâmetros de configuração na tabela 3.13.

Tabela 3.13 Parâmetros de configuração do conjunto de dados sintéticos 3

	Classe 1		Classe 2	
	Média	Variância	Média	Variância
x_1	3	0.1	3	0.1
x_2	3	0.1	3	0.1
x_3	3	0.1	-1	0.1

Tabela 3.14 Matriz de confusão obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 3 com 2 grupos de variáveis

Classe	P_1	P_2	Somatório
1	100	0	100
2	0	100	100
Somatório	100	100	200

A aplicação do algoritmo BCKM coloca a variável x_3 em H_2 e as variáveis x_1 e x_2 em H_1, já os indivíduos são alocados em clusters de acordo com suas classes à priori como pode ser percebido pelo índice OERC de 0 na tabela 3.17 que é igual a execução para o algoritmo KM. Então, podemos perceber que a inclusão individual de x_3 no grupo H_2 é devido a termos valores variando entre -2 e 2 que é diferente dos valores que x_1 e x_2 têm associadas aos indivíduos, sendo para ambas entre 2 e 4, assim apesar da homogeneidade dos indivíduos, em geral, ser menor em x_3, em relação a x_1 e x_2, a sua característica de definição dos grupos de indivíduos torna as variável x_3 mais homogênea em cada parcela de grupos de indivíduos formados, que pode ser visto na matriz do critério de a equação localizada na tabela 3.18, que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e é sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis, que relaciona a cada bloco mostrado na tabela 3.16, onde podemos ver a similaridade dos indivíduos no mesmo bloco e a dissimilaridade dos que estão em blocos diferentes.

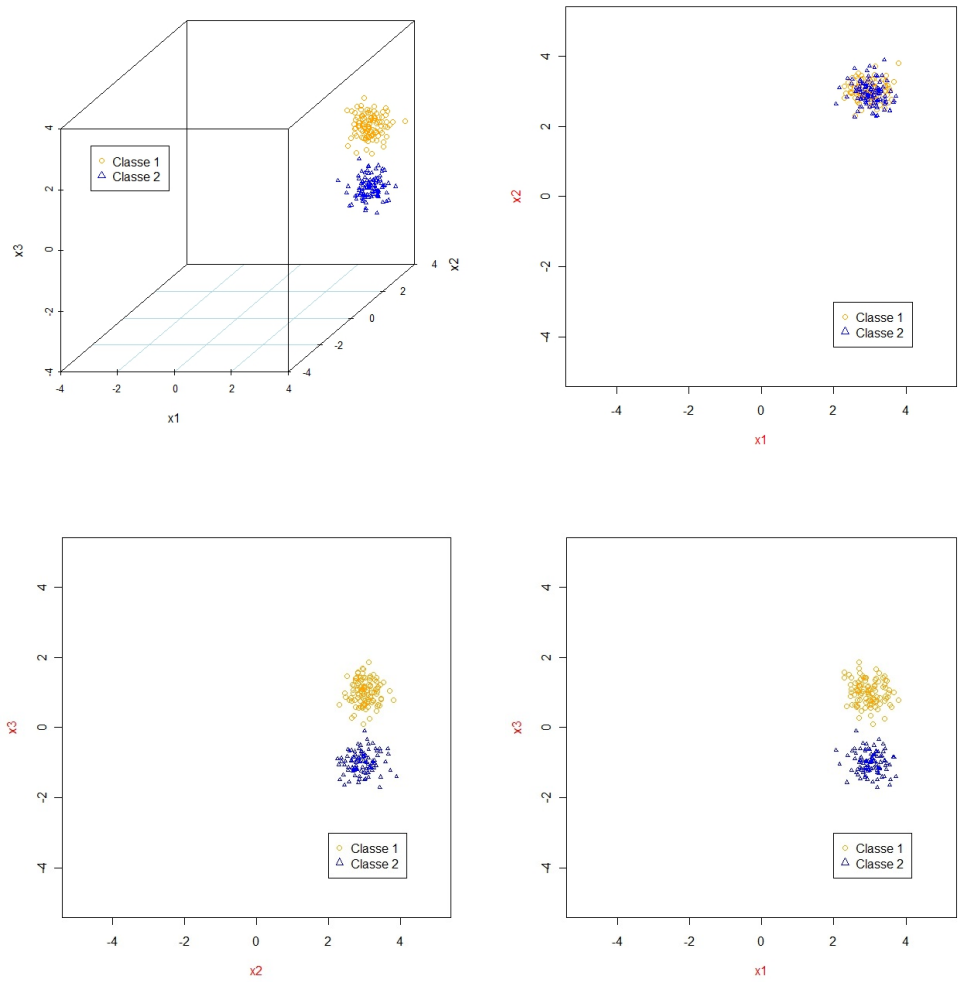

Figura 3.3 Conjunto de dados Sintético 3

Tabela 3.15 Protótipos obtidos para cada um dos blocos para o conjunto de dados sintéticos 3

	H_1	H_2
P_1	2.9923	0.9985
P_2	2.9777	-1.0030

Tabela 3.16 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados sintéticos 3

		H_1		H_2
		x_1	x_2	x_3
	e_1	2.9434	2.9915	1.1643
	e_2	2.7065	3.0584	1.0136
	e_3	2.7521	2.8445	0.9288
	e_4	2.4780	2.9371	1.0817
P_1	e_5	2.8710	2.9559	0.5635
	e_6	2.6536	3.3538	0.6100
	e_7	2.4649	2.9876	0.8861
	e_8	3.7956	3.7981	0.7807
	e_9	3.4413	3.4556	0.8131
	e_{10}	3.0569	3.3407	0.9634
	e_{101}	2.7344	3.3105	-0.6075
	e_{102}	2.3873	3.3586	-0.8078
	e_{103}	3.0733	3.2372	-0.9841
	e_{104}	3.3666	3.1138	-0.4774
P_2	e_{105}	3.1722	2.2814	-1.0510
	e_{106}	3.3671	2.7307	-0.6774
	e_{107}	3.2643	3.0631	-1.1673
	e_{108}	3.1621	3.4470	-0.6902
	e_{109}	3.2319	2.7411	-1.1527
	e_{110}	3.4017	2.4710	-0.9518

Tabela 3.17 Performance dos algoritmos aplicados ao conjunto de dados sintéticos 3

	OERC	CR
BCKM	0	1
KM	0	1

Tabela 3.18 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 3

	H_1	H_2
P_1	18.3199	**11.7345**
P_2	21.1922	**9.0305**

3.2.1.4 Conjunto de dados sintéticos 4

Esse conjunto de dados sintéticos foi gerado com duas classes que possuem 100 indivíduos cada, sendo os indivíduos descritos por 4 variáveis, onde as variáveis x_1 e x_2 possuem valores

entre 2 e 4, e as variáveis x_3 e x_4 possuem valores entre -2 e 2, assim os valores de x_1 e x_2 são diferentes dos de x_3 e x_4 (figura 3.4), esses valores foram gerados de acordo com a configuração descrita na tabela 3.19.

Figura 3.4 Conjunto de dados Sintético 4

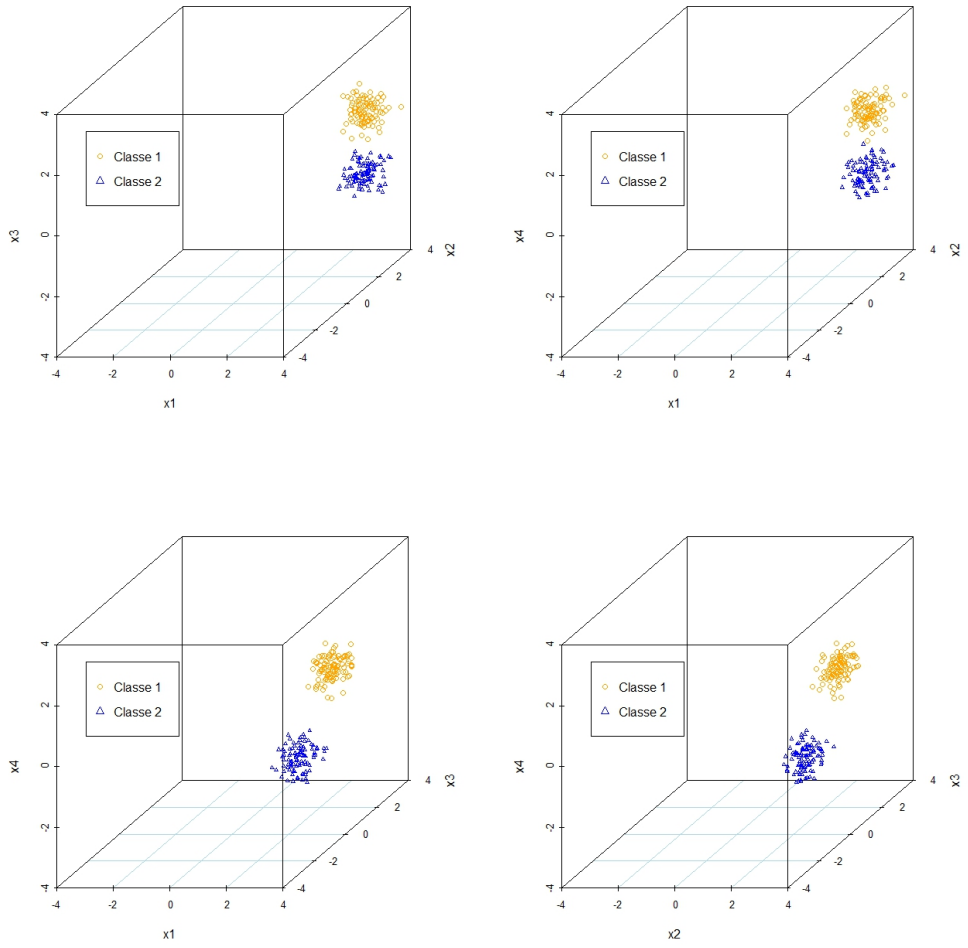

Tabela 3.19 Parâmetros de configuração do conjunto de dados sintéticos 4

	Classe 1		Classe 2	
	Média	Variância	Média	Variância
x_1	3	0.1	3	0.1
x_2	3	0.1	3	0.1
x_3	3	0.1	-1	0.1
x_4	3	0.1	-1	0.1

Tabela 3.20 Performance dos algoritmos aplicados ao conjunto de dados sintéticos 4

	OERC	CR
BCKM	0	1
KM	0	1

Tabela 3.21 Matriz de confusão obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 4 com 2 grupos de variáveis

Classe	P_1	P_2	Somatório
1	100	0	100
2	0	100	100
Somatório	100	100	200

Tabela 3.22 Protótipos obtidos para cada um dos blocos para o conjunto de dados sintéticos 4

	H_1	H_2
P_1	2.9923	1.0065
P_2	2.9688	-1.0316

Tabela 3.23 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados sintéticos 4

	H_1	H_2
P_1	**18.3199**	21.8256
P_2	**20.0297**	22.2320

Tabela 3.24 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados sintéticos 4

		H_1		H_2	
		x_1	x_2	x_3	x_4
	e_1	2.9434	2.9915	1.1643	0.7344
	e_2	2.7065	3.0584	1.0136	0.3873
	e_3	2.7521	2.8445	0.9288	1.0733
	e_4	2.4780	2.9371	1.0817	1.3666
P_1	e_5	2.8710	2.9559	0.5635	1.1722
	e_6	2.6536	3.3538	0.6100	1.3671
	e_7	2.4649	2.9876	0.8861	1.2643
	e_8	3.7956	3.7981	0.7807	1.1621
	e_9	3.4413	3.4556	0.8131	1.2319
	e_{10}	3.0569	3.3407	0.9634	1.4017
	e_{101}	3.3105	3.3925	-0.6878	-1.0865
	e_{102}	3.3586	3.1923	-1.5221	-0.7462
	e_{103}	3.2372	3.0159	-0.7006	-0.7709
	e_{104}	3.1138	3.5226	-0.9787	-0.7925
P_2	e_{105}	2.2814	2.9490	-0.8788	-0.8245
	e_{106}	2.7307	3.3226	-0.9467	-0.6910
	e_{107}	3.0631	2.8327	-1.2223	-1.1910
	e_{108}	3.4470	3.3098	-1.2951	-1.3432
	e_{109}	2.7411	2.8474	-1.0805	-1.7821
	e_{110}	2.4710	3.0482	-1.1652	-1.4208

Como resultado as variáveis x_1 e x_2 são alocadas no mesmo grupo H_1 enquanto as variáveis x_3 e x_4 ficam em H_2, e os grupos de indivíduos formados são de acordo com as classes à priori, como indica o índice OERC de 0 na tabela 3.20, que é igual para os algoritmos KM e BCKM. Assim, é notado pela matriz de blocos, tabela 3.24, que as variáveis com valores similares ficaram no mesmo grupo de variáveis, enquanto os indivíduos foram separados de acordo com os valores deles nas variáveis x_3 e x_4, que distingue bem dois grupos, assim a homogeneidade para cada bloco pode ser analisada pela matriz do critério de adequação na tabela 3.23, que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e é sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis.

3.2.2 Conjunto de dados reais

De forma a comparar a performance, através dos índices OERC e CR, os algoritmos BCKM e KM são submetidos a 3 experimentos que se diferenciam pela base de dados a ser utilizada, então as bases utilizadas em cada experimento são: Iris Plant, Thyroid gland e Wine que podem ser obtidas no repositório de bases de dados UCI Machine Learning Repository, uma rápida descrição delas pode ser vista na tabela 3.25. Cada experimento é feito para configurações diferentes de clusters de variáveis e os clusters de indivíduos são iguais a quantidade de classes a priori da base de dados.

Tabela 3.25 Descrição geral das bases de dados reais quatitativas utilizadas

	Iris Plant	Thyroid Gland	Wine
Qtd. Padrões	150	215	178
Qtd. Variáveis	4	5	13
Qtd. Classes à priori	3	3	3

3.2.2.1 Iris Plant

A base de dados Iris possui 150 padrões que são descritos por 4 variáveis reais: (1) sepal length, (2) sepal width, (3) petal length e (4) petal width. É dividida em 3 classes a priori, que são elas *Iris setosa*, *Iris versicolour* e *Iris virginica*, com 50 padrões cada. Assim, aplicamos o algoritmo para 3 clusters de indivíduos e para 2 clusters de variáveis para o primeiro teste, depois 3 clusters de variáveis para um segundo teste. Podemos ver através da figura 3.5 que há uma grande variação nos valores possíveis dos indivíduos na variável x_3, quando comparado às outras variáveis.

Tabela 3.26 Performance dos algoritmos aplicados ao conjunto de dados Iris Plant

	Qtd. Clusters de variáveis	OERC	CR
BCKM	1	0.1466	0.6403
BCKM	2	0.1466	0.6403
BCKM	3	0.1466	0.6403
BCKM	4	0.1066	0.7302
KM	-	0.1066	0.7302

Tabela 3.27 Protótipos obtidos para cada um dos blocos para o conjunto de dados Iris Plant com 2 clusters de variáveis

	H_1	H_2
P_1	3.3018	1.8185
P_2	5.2172	2.1155
P_3	6.2960	2.5803

Tabela 3.28 Protótipos obtidos para cada um dos blocos para o conjunto de dados Iris Plant com 3 clusters de variáveis

	H_1	H_2	H_3
P_1	4.4039	2.0710	6.8737
P_2	2.4667	0.3018	5.0056
P_3	3.6276	1.4620	5.9483

Figura 3.5 Conjunto de dados Iris Plant

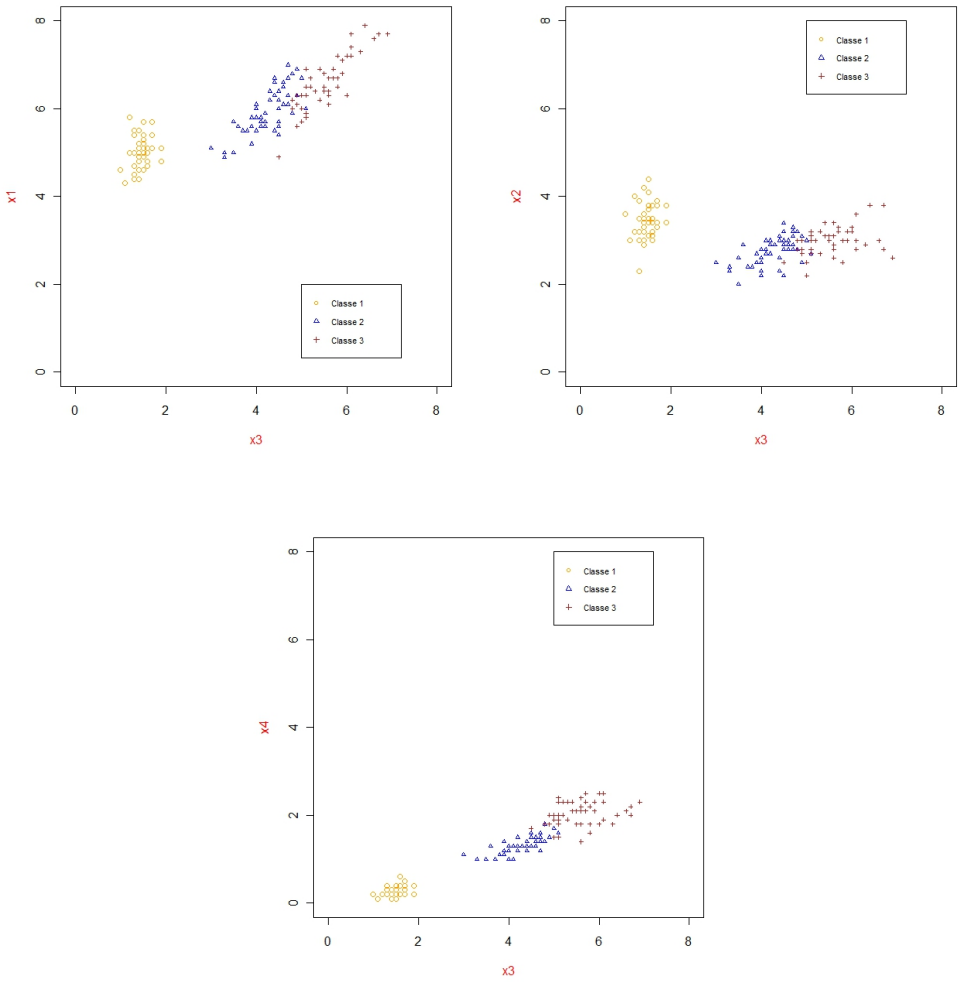

Tabela 3.31 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Iris Plant com 2 grupos de variáveis

	H_1	H_2
P_1	333.3395	**263.1430**
P_2	82.2255	**58.4120**
P_3	43.7088	**25.5004**

Tabela 3.29 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados Iris Plant com 2 grupos de variáveis

		H_1		H_2	
		x_1	x_3	x_2	x_4
	e_2	6.9	4.9	3.1	1.5
	e_4	6.7	5.7	3.3	2.5
	e_8	6.3	5.6	3.4	2.4
	e_{11}	7.7	6.1	3.0	2.3
P_1	e_{19}	7.0	4.7	3.2	1.4
	e_{26}	6.7	5.6	3.1	2.4
	e_{27}	6.4	5.3	2.7	1.9
	e_{29}	7.7	6.7	3.8	2.2
	e_{32}	7.9	6.4	3.8	2.0
	e_{35}	6.8	5.9	3.2	2.3
	e_3	5.1	1.6	3.8	0.2
	e_5	5.2	1.5	4.1	0.1
	e_7	5.1	3.0	2.5	1.1
	e_{12}	4.8	1.6	3.1	0.2
P_2	e_{15}	5.5	1.4	4.2	0.2
	e_{16}	4.4	1.4	2.9	0.2
	e_{18}	5.1	1.4	3.5	0.3
	e_{20}	4.8	1.4	3.0	0.1
	e_{23}	4.6	1.0	3.6	0.2
	e_{25}	4.9	1.5	3.1	0.1
	e_1	6.3	4.9	2.5	1.5
	e_6	5.9	4.2	3.0	1.5
	e_9	6.1	5.6	2.6	1.4
	e_{10}	6.0	4.8	3.0	1.8
P_3	e_{13}	6.2	4.3	2.9	1.3
	e_{14}	6.6	4.4	3.0	1.4
	e_{17}	5.5	4.0	2.3	1.3
	e_{21}	5.6	4.2	2.7	1.3
	e_{22}	5.7	4.5	2.8	1.3
	e_{24}	5.7	4.2	3.0	1.2

Tabela 3.30 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados Iris Plant com 3 grupos de variáveis

		H_1		H_2	H_3
		x_2	x_3	x_4	x_1
	e_2	3.1	4.9	1.5	6.9
	e_4	3.3	5.7	2.5	6.7
	e_8	3.4	5.6	2.4	6.3
	e_{11}	3.0	6.1	2.3	7.7
P_1	e_{19}	3.2	4.7	1.4	7.0
	e_{26}	3.1	5.6	2.4	6.7
	e_{27}	2.7	5.3	1.9	6.4
	e_{29}	3.8	6.7	2.2	7.7
	e_{32}	3.8	6.4	2.0	7.9
	e_{35}	3.2	5.9	2.3	6.8
	e_3	3.8	1.6	0.2	5.1
	e_5	4.1	1.5	0.1	5.2
	e_7	2.5	3.0	1.1	5.1
	e_{12}	3.1	1.6	0.2	4.8
P_2	e_{15}	4.2	1.4	0.2	5.5
	e_{16}	2.9	1.4	0.2	4.4
	e_{18}	3.5	1.4	0.3	5.1
	e_{20}	3.0	1.4	0.1	4.8
	e_{23}	3.6	1.0	0.2	4.6
	e_{25}	3.1	1.5	0.1	4.9
	e_1	2.5	4.9	1.5	6.3
	e_6	3.0	4.2	1.5	5.9
	e_9	2.6	5.6	1.4	6.1
	e_{10}	3.0	4.8	1.8	6.0
P_2	e_{13}	2.9	4.3	1.3	6.2
	e_{14}	3.0	4.4	1.4	6.6
	e_{17}	2.3	4.0	1.3	5.5
	e_{21}	2.7	4.2	1.3	5.6
	e_{22}	2.8	4.5	1.3	5.7
	e_{24}	3.0	4.2	1.2	5.7

Tabela 3.32 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Iris Plant com 3 grupos de variáveis

	H_1	H_2	H_3
P_1	144.0890	**2.8981**	8.4737
P_2	107.1000	**2.8298**	6.1083
P_3	101.2515	**4.6765**	8.7048

Após a execução do algoritmo BCKM para 2 grupos de variáveis temos x_1 e x_3 alocados ao grupo H_1 e x_2 e x_4 em H_2. Assim, analisando a variável x_3, podemos perceber através da figura 3.5 e da tabela 3.29 que descreve os blocos obtidos, que os grupos de indivíduos são bem separados em x_3, mas o grupo de variáveis ao qual ela se insere traz um índice de adequação (tabela 3.31) para seus blocos maior que os do grupo H_2, dada a grande heterogeneidade dos valores dos indivíduos nessa variável, e consequentemente participa do bloco com pior adequação(P_1,H_1), onde os valores das variáveis para os indivíduos nesse bloco são bastante diferentes.

Quando aumentamos o número de clusters de variáveis para 3 é notado que as variáveis x_2 e x_3 são alocadas ao mesmo grupo H_1, enquanto x_4 fica no grupo H_2 e x_1 fica no em H_3. Analisando o índice de adequação da configuração com 3 clusters de variáveis na tabela 3.32, que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e é sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis, vemos que o grupo H_1 é bem heterogêneo, dada a inclusão de x_3, enquanto os outros dois grupos que incluem as outras duas variáveis trazem índices de adequação mais baixos e consequentemente maior homogeneidade, quando comparado a H_1.

Essa situação de x_2 e x_3 estarem em grupos diferentes para o primeiro teste, mas ficarem no mesmo grupo para o segundo teste é válida, no sentido que elas tem valores mais similares entre si que com as outras variáveis e x_1 e x_4 que são muito dissimilares entre si, mas x_1 tem uma maior proximidade com a variável x_3 e x_4 tem com x_2, sendo assim é adequado para o primeiro teste termos um grupo com x_1 e x_3 e outro com x_2 e x_4, já que temos apenas dois grupos de variáveis disponiveis.

Comparando a performance entre os modelos BCKM e KM pelos índices da tabela 3.26 vemos que ela é um pouco degradada tanto para configuração com 2 como para configuração com 3 grupos de variáveis.

3.2.2.2 Thyroid Gland

A base de dados thyroid gland possui 215 instâncias divididas desigualmente entre as 3 classes que descrevem o estado desta glândula: normal (com 150 instâncias), hyperthyroidism (com 30 instâncias) e hypothyroidism (com 30 instâncias). Cada instância é descrita por 5 variáveis de tipo real que são: (1) T3-resin uptake test, (2) total serum thyroxin, (3) total serum triiodothyronine, (4) basal thyroidstimulating hormone (TSH) e (5) maximal absolute difference of TSH. Foram feitos 3 testes, de forma que o primeiro possui 3 grupos de variáveis, o segundo possui 4 grupos de variáveis e o terceiro possui 5 grupos de variáveis, e todos possuem 3 grupos de indivíduos.

Tabela 3.33 Performance dos algoritmos aplicados ao conjunto de dados Thyroid Gland

	Qtd. Clusters de variáveis	OERC	CR
BCKM	3	0.1581	0.5437
BCKM	4	0.1395	0.5790
BCKM	5	0.1395	0.5790
KM	-	0.1395	0.5790

Tabela 3.34 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Thyroid Gland com 3 grupos de variáveis

	H_1	H_2	H_3
P_1	**1137.9750**	6158.4500	1485.2750
P_2	866.7200	2788.8550	**473.0250**
P_3	**884.9700**	1996.6200	16636.7500

Tabela 3.35 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Thyroid Gland com 4 grupos de variáveis

	H_1	H_2	H_3	H_4
P_1	243.0700	3933.8950	1367.8750	**234.3150**
P_2	9555.0000	1605.6500	63.3600	**6.1850**
P_3	1005.9650	9480.7000	1335.6400	**55.5350**

Tabela 3.36 Protótipos obtidos para cada um dos blocos para o conjunto de dados Thyroid com 3 clusters de variáveis

	H_1	H_2	H_3
P_1	9.3098	109.7778	1.9159
P_2	16.8000	86.5714	1.9773
P_3	6.2705	127.7352	8.9931

Tabela 3.37 Protótipos obtidos para cada um dos blocos para o conjunto de dados Thyroid com 4 clusters de variáveis

	H_1	H_2	H_3	H_4
P_1	1.0564	89.9487	14.7179	3.6461
P_2	18.3391	125.5652	3.2000	0.9739
P_3	1.9477	112.2026	9.5451	1.8052

Tabela 3.38 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados Thyroid Gland com 3 grupos de variáveis

		H_1	H_2	H_3		
		x_2	x_1	x_3	x_4	x_5
	e_1	10.1	107.0	2.2	0.9	2.7
	e_2	9.9	113.0	3.1	2.0	5.9
	e_4	5.3	109.0	1.6	1.4	1.5
P_1	e_5	7.3	105.0	1.5	1.5	-0.1
	e_6	6.1	105.0	2.1	1.4	7.0
	e_7	10.4	110.0	1.6	1.6	2.7
	e_8	9.9	114.0	2.4	1.5	5.7
	e_9	9.4	106.0	2.2	1.5	0.0
	e_{10}	13.0	107.0	1.1	0.9	3.1
	e_{11}	4.2	106.0	1.2	1.6	1.4
	e_{34}	8.1	90.0	1.6	1.4	1.1
	e_{57}	8.9	93.0	1.5	0.8	2.7
	e_{71}	9.4	96.0	1.5	1.0	3.1
	e_{82}	8.0	91.0	1.7	2.1	4.6
P_2	e_{145}	7.5	94.0	1.2	1.3	4.4
	e_{154}	25.3	65.0	5.8	1.3	0.2
	e_{155}	24.1	88.0	5.5	0.8	0.1
	e_{156}	18.2	65.0	10.0	1.3	0.1
	e_{159}	23.3	67.0	7.4	1.8	-0.6
	e_{160}	11.1	95.0	2.7	1.6	-0.3
	e_3	12.9	127.0	2.4	1.4	0.6
	e_{39}	9.5	130.0	1.7	0.4	3.2
	e_{43}	11.9	129.0	2.7	1.2	3.5
	e_{45}	8.1	123.0	2.3	1.0	5.1
P_3	e_{73}	9.7	133.0	2.9	0.8	1.9
	e_{74}	9.4	126.0	2.3	1.0	4.0
	e_{129}	10.0	130.0	1.6	0.9	4.6
	e_{138}	7.7	127.0	1.8	1.9	6.4
	e_{146}	10.4	126.0	1.7	1.2	3.5
	e_{151}	16.4	139.0	3.8	1.1	-0.2

Tabela 3.39 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados Thyroid Gland com 4 grupos de variáveis

		H_1		H_2	H_3	H_4
		x_4	x_5	x_1	x_2	x_3
	e_{34}	1.4	1.1	90.0	8.1	1.6
	e_{40}	0.9	1.9	100.0	10.5	2.4
	e_{49}	1.3	-0.2	100.0	9.5	2.5
	e_{51}	0.7	-0.3	100.0	11.3	2.5
P_1	e_{57}	0.8	2.7	93.0	8.9	1.5
	e_{68}	1.2	0.9	97.0	7.8	1.3
	e_{71}	1.0	3.1	96.0	9.4	1.5
	e_{82}	2.1	4.6	91.0	8.0	1.7
	e_{84}	1.9	-0.3	98.0	9.1	1.4
	e_{94}	1.6	6.0	98.0	8.6	1.6
	e_{186}	16.5	9.5	125.0	2.3	0.9
	e_{187}	10.4	38.6	120.0	6.8	2.1
	e_{190}	23.0	5.7	119.0	3.8	1.1
	e_{191}	9.2	14.4	141.0	5.6	1.8
P_2	e_{192}	12.5	2.9	129.0	1.5	0.6
	e_{193}	11.6	48.8	118.0	3.6	1.5
	e_{194}	18.5	24.0	120.0	1.9	0.7
	e_{195}	56.4	21.6	119.0	0.8	0.7
	e_{196}	13.7	56.3	123.0	5.6	1.1
	e_{198}	12.2	8.8	126.0	0.5	0.2
	e_1	0.9	2.7	107.0	10.1	2.2
	e_2	2.0	5.9	113.0	9.9	3.1
	e_3	1.4	0.6	127.0	12.9	2.4
	e_4	1.4	1.5	109.0	5.3	1.6
P_3	e_5	1.5	-0.1	105.0	7.3	1.5
	e_6	1.4	7.0	105.0	6.1	2.1
	e_7	1.6	2.7	110.0	10.4	1.6
	e_8	1.5	5.7	114.0	9.9	2.4
	e_9	1.5	0.0	106.0	9.4	2.2
	e_{10}	0.9	3.1	107.0	13.0	1.1

Tabela 3.40 Descrição da alocação das variáveis nos grupos de variável em cada uma das configurações testadas

	Qtd. Clusters de variáveis		
	3 Clusters	4 Clusters	5 Clusters
H_1	x_3, x_4, x_5	x_4, x_5	x_4
H_2	x_1	x_1	x_1
H_3	x_2	x_2	x_2
H_4	-	x_3	x_3
H_5	-	-	x_5

Com o aumento da quantidade de clusters de variáveis disponíveis as variáveis vão se adaptando a quantidade disponível de acordo com a similaridade entre as variáveis, de forma que fica cada uma em um grupo para o teste em que a quantidade de grupos de variáveis é igual a quantidade de variáveis como podemos ver na tabela 3.40, que descreve a alocação das variáveis nos grupos de variáveis para cada configuração de clusters de variáveis testada. Além disso, comparando a performance dos modelos BCKM e KM percebemos, pela tabela 3.33, que com o aumento do número de grupos o critério de adequação geral e o índice OERC são reduzidos, chegando a serem iguais quando temos uma quantidade de grupos de variáveis idêntica ao número de variáveis, que faz sentido já que estamos retirando a restrição de variáveis estarem interligadas através de determinado grupo de variáveis, que causa maior erro quando uma das variáveis nesse grupo for muito heterogênea.

Os blocos com os 10 primeiros padrões de cada cluster de variável para as configurações com 3 e 4 grupos de variáveis podem ser vistos pelas tabelas 3.38 e 3.39, respectivamente, podendo assim ser percebido a proximidade dos valores dos indivíduos em determinado bloco, que fica mais aparente quanto menor for o índice de adequação para aquele bloco, tabelas 3.34 e 3.35 destacam em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis.

3.2.2.3 Wine

Com 178 padrões a base de dados wine é distribuída em 3 classes com 59, 71 e 48 padrões, respectivamente, e descreve vinhos oriundos da mesma região da Itália, mas derivados de cultivos diferente. A base de dados é descrita pelas 13 variáveis reais seguintes: (1) alcohol, (2) malic acid, (3) ash, (4) alkalinity of ash, (5) magnesium, (6) total phenols, (7) flavonoids, (8) non-flavonoid phenols, (9) proanthocyanins, (10) colour intensity, (11) hue, (12) OD280/OD315 of diluted wines e (13) proline. Foram analisadas 3 configurações para o modelo BCKM onde todas têm 3 clusters de indivíduos mas que se diferenciam no número de clusters de variáveis que são 3 clusters para primeira, 5 cluster para a segunda e 8 clusters para terceira e todas foram comparadas com o modelo KM.

O resultado apresentado na tabela 3.42 mostram que a performance do modelo KM se mantém no modelo BCKM para todos os teste. Pela tabela 3.41 vemos que no agrupamento por 5 cluster as variáveis se concentram em um grupo já que os outros grupos tem variáveis com valores muito diferentes das outras, quando aumentamos para 8 clusters de variáveis as variáveis x_7 e x_9 vão para um grupo enquanto as variáveis x_8 e x_{11} vão para outro mostrando

assim que os grupos se dividem quanto maior for a proximidade entre os valores dos indivíduos naquelas variáveis de acordo com a disponibilidade de grupos para se ter uma divisão mais adequada. Na tabela 3.45 podemos ver a descrição dos blocos para a configuração com 5 clusters de variáveis e na tabela 3.43 os critérios de adequação para cada bloco, sendo em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis.

Tabela 3.41 Descrição da alocação das variáveis nos grupos de variável em cada um dos testes na base de dados Wine

	Qtd. Clusters de variáveis		
	3 Clusters	5 Clusters	8 Clusters
H_1	$x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}$	$x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}$	x_2, x_3, x_6, x_{12}
H_2	x_{13}	x_{13}	x_{13}
H_3	x_5	x_5	x_5
H_4	-	x_1	x_1
H_5	-	x_4	x_4
H_6	-	-	x_7, x_9
H_7	-	-	x_8, x_{11}
H_8	-	-	x_{10}

Tabela 3.42 Performance dos algoritmos aplicados ao conjunto de dados Wine

	Qtd. Clusters de variáveis	OERC	CR
BCKM	3	0.2977	0.3711
BCKM	5	0.2977	0.3711
BCKM	8	0.2977	0.3711
KM	-	0.2977	0.3711

Tabela 3.43 Matriz do critério de adequação para cada bloco (J_{kh}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Wine com 5 clusters de variáveis

	H_1	H_2	H_3	H_4	H_5
P_1	1354680.0000	1001.3950	**9.9177**	5725.7500	413.7245
P_2	434590.5000	1187.8150	**30.2521**	7297.6500	681.7250
P_3	548585.0000	1784.0600	**29.6429**	16824.9000	460.0540

Tabela 3.44 Protótipos obtidos para cada um dos blocos para o conjunto de dados Wine com 5 clusters de variáveis

	H_1	H_2	H_3	H_4	H_5
P_1	1195.1489	2.4757	13.8044	105.5106	17.0234
P_2	458.2319	1.9970	12.5167	92.3478	20.8231
P_3	728.3387	2.1554	12.9298	103.5968	19.8903

Tabela 3.45 Matriz de blocos para os 10 primeiros indivíduos de cada grupo de indivíduos para o conjunto de dados Wine com 5 clusters de variáveis

		H_1	H_2									H_3	H_4	H_5
		x_{13}	x_2	x_3	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_1	x_5	x_4
	e_1	1065.0	1.7	2.4	2.8	3.1	0.3	2.3	5.6	1.0	3.9	14.2	127.0	15.6
	e_2	1050.0	1.8	2.1	2.7	2.8	0.3	1.3	4.4	1.1	3.4	13.2	100.0	11.2
	e_3	1185.0	2.4	2.7	2.8	3.2	0.3	2.8	5.7	1.0	3.2	13.2	101.0	18.6
	e_4	1480.0	2.0	2.5	3.9	3.5	0.2	2.2	7.8	0.9	3.5	14.4	113.0	16.8
P_1	e_6	1450.0	1.8	2.5	3.3	3.4	0.3	2.0	6.8	1.1	2.9	14.2	112.0	15.2
	e_7	1290.0	1.9	2.5	2.5	2.5	0.3	2.0	5.3	1.0	3.6	14.4	96.0	14.6
	e_8	1295.0	2.2	2.6	2.6	2.5	0.3	1.3	5.1	1.1	3.6	14.1	121.0	17.6
	e_9	1045.0	1.6	2.2	2.8	3.0	0.3	2.0	5.2	1.1	2.9	14.8	97.0	14.0
	e_{10}	1045.0	1.4	2.3	3.0	3.2	0.2	1.9	7.2	1.0	3.6	13.9	98.0	16.0
	e_{11}	1510.0	2.2	2.3	3.0	3.3	0.2	2.4	5.8	1.3	3.2	14.1	105.0	18.0
	e_{60}	520.0	0.9	1.4	2.0	0.6	0.3	0.4	2.0	1.1	1.8	12.4	88.0	10.6
	e_{62}	450.0	1.4	2.0	2.0	1.4	0.5	0.6	5.8	1.0	1.6	12.6	100.0	16.8
	e_{64}	420.0	1.1	2.2	3.5	3.1	0.2	1.9	4.5	1.2	2.9	12.4	87.0	19.0
	e_{65}	355.0	1.5	2.5	1.9	1.8	0.5	1.0	3.0	1.5	2.2	12.2	104.0	19.0
P_2	e_{67}	502.0	1.0	1.7	3.0	3.2	0.3	2.3	5.3	1.1	3.2	13.1	78.0	15.0
	e_{68}	510.0	1.2	1.9	2.1	2.0	0.3	1.0	4.7	1.1	3.5	12.4	78.0	19.6
	e_{72}	410.0	1.5	2.7	3.0	2.9	0.2	1.9	3.4	1.4	3.2	13.9	86.0	25.0
	e_{73}	472.0	1.7	2.2	1.9	1.8	0.3	1.0	3.7	1.0	2.8	13.5	87.0	24.0
	e_{76}	428.0	1.9	1.9	1.6	1.6	0.3	1.2	3.8	1.2	2.1	11.7	97.0	16.0
	e_{77}	392.0	0.9	1.7	2.0	2.0	0.2	1.5	4.6	1.2	2.5	13.0	86.0	16.0
	e_5	735.0	2.6	2.9	2.8	2.7	0.4	1.8	4.3	1.0	2.9	13.2	118.0	21.0
	e_{20}	845.0	3.1	2.6	2.7	3.0	0.2	1.7	5.1	1.0	3.4	13.6	116.0	15.2
	e_{21}	780.0	1.6	2.3	3.0	3.2	0.2	2.1	5.7	1.1	3.7	14.1	126.0	16.0
	e_{22}	770.0	3.8	2.7	2.4	2.4	0.3	2.0	4.5	1.0	3.5	12.9	102.0	18.6
P_3	e_{25}	845.0	1.8	2.6	2.5	2.6	0.3	1.7	3.5	1.1	3.8	13.5	96.0	20.0
	e_{26}	830.0	2.1	3.2	2.6	2.7	0.5	1.9	3.6	1.1	3.2	13.1	124.0	25.0
	e_{29}	915.0	1.9	2.8	3.0	3.0	0.4	1.8	4.5	1.3	3.4	13.9	107.0	19.4
	e_{36}	920.0	1.8	2.4	2.7	3.0	0.3	1.9	5.1	1.0	3.5	13.5	100.0	20.5
	e_{37}	880.0	1.6	2.8	2.6	2.7	0.3	1.4	4.6	1.1	2.8	13.3	110.0	15.5
	e_{40}	760.0	4.0	2.5	3.0	3.0	0.2	2.1	5.1	0.9	3.5	14.2	128.0	13.2

3.2.3 Conclusões parciais

Nessa seção fizemos uma análise do algoritmo de Block clustering K-means para dados quantitativo, assim a partir dos dados simulados vimos que os blocos se dividem de acordo com a homogeneidade dos valores dos indivíduos em cada variável, formando-se grupos de indivíduos que influenciam nos grupos de variáveis e grupos de variáveis que influenciam na formação dos grupos de indivíduos. O particionamento do grupo de variáveis é feito de acordo com a homogeneidade dos valores das variáveis, sendo que quanto mais homogêneas os valores das variáveis em determinado grupo de indivíduos os grupos variáveis são formados mais precisamente, no caso, como valores mais semelhantes, e quanto menos homogêneos forem os valores das variáveis nos grupos de indivíduos a formação dos grupos será menos precisa, tendo valores mais diferentes em cada bloco. Para a aplicação em dados reais do algoritmo BCKM comparamos com o algoritmo KM e percebemos que apesar de termos uma informação a mais, a inclusão da partição de variáveis, o desempenho não é degradado a partir de determinada quantidade de grupos de variáveis escolhida.

CAPÍTULO 4

Block Clustering K-means para Variáveis do Tipo Intervalo

Nessa seção iremos apresentar o algoritmo de Block Clustering K-means(BCKM) para Variáveis do Tipo Intervalo e fazer alguns experimentos em bases de dados reais.

Tal algoritmo tem como objetivo fornecer a partição de um conjunto de itens em K grupos $P_1,\ldots,P_K$ de itens e as variáveis correspondentes de cada item em Q grupos $H_1,\ldots,H_Q$ de variáveis e deve fornecer também seus centróides (protótipos) correspondentes $g_{kq} = [\alpha_{kq},\beta_{kq}](k = 1,\ldots,K)(q = 1,\ldots,Q)$ de forma a otimizar (minimizar) o critério de adequação J, que mede o ajuste entre os grupos e os protótipos. A seguir podemos ver a descrição do critério J:

$$J = \sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})[(a_{ij}-\alpha_{kq})^2 + (b_{ij}-\beta_{kq})^2] \tag{4.1}$$

onde $U_{ik}(i = 0,\ldots,n)(k = 0,\ldots,K)$ é uma matriz que indica a pertinência do indivíduo e_i ao grupo P_k, assim $U_{ik} = 1$ para $e_i \in P_k$ e $U_{ik'} = 0$ para qualquer outro grupo $P_{k'}(k' = 1,\ldots,K)$ e $k' \neq k$, já $V_{jq}(j = 0,\ldots,p)(q = 0,\ldots,Q)$ indica a pertinência da variável x_j ao grupo de variável H_q, então $V_{jq} = 1$ para $x_j \in H_q$ e $V_{jq'} = 0$ para qualquer outro grupo $H_{q'}(q' = 1,\ldots,Q)$ e $q' \neq q$. Temos a_{ij} como o valor de mínimo e b_{ij} como o valor de máximo do indivíduo $e_i(i = 1,\ldots,n)$ para a variável $x_j(j = 1,\ldots,p)$.

Os protótipos são computados da seguinte maneira:

$$\alpha_{kq} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})a_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})} \quad e \quad \beta_{kq} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})b_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})} \tag{4.2}$$

É possível notar que $\alpha_{kq} \leq \beta_{kq}$, já que $a_{ij} \leq b_{ij}$.

Os indivíduos devem ser alocados em grupos de indivíduos de forma que o critério J seja minimizado, assim o indivíduo e_i deve ser alocado ao grupo k seguindo o seguinte critério:

$$k = \operatorname*{argmin}_{1\leq l\leq K}\sum_{q=1}^{Q}\sum_{j=1}^{p}(V_{jq})[(a_{ij}-\alpha_{lq})^2 + (b_{ij}-\beta_{lq})^2] \tag{4.3}$$

As variáveis devem ser alocadas em grupos de variáveis de forma que o critério J seja minimizado, assim a variável x_j deve ser alocado ao grupo q seguindo o seguinte critério:

$$q = \underset{1\leq l\leq Q}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{n}(U_{ik})[(a_{ij}-\alpha_{kl})^2+(b_{ij}-\beta_{kl})^2] \tag{4.4}$$

4.1 Descrição do algoritmo

Para execução do algoritmo Block Clustering K-Means(BCKM) para variáveis de tipo intervalo devemos seguir os seguintes passos:

1) *Inicialização:*

   Definir o número K de grupos de indivíduos e o número H de grupos de variáveis.

   Aleatoriamente construir a partição inicial $P^{(0)} = (P_1^{(0)},\ldots,P_k^{(0)},\ldots,P_K^{(0)})$ de indivíduos, U_{ik} deve estar de acordo com essa partição, e a partição inicial $H^{(0)} = (H_1^{(0)},\ldots,H_q^{(0)},\ldots,H_Q^{(0)})$ de variáveis e V_{jq} corresponder a essa partição.

2) *Passo 1: Calcular protótipos*

   Computar os componentes $g_{kq} = [\alpha_{kq},\beta_{kq}]$ dos protótipos $g_{kq}(k = 1,\ldots,K)(q = 1,\ldots,Q)$ de acordo com a equação (4.2):

$$\alpha_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})a_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})} \quad e \quad \beta_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})b_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})}$$

3) *Passo 2: Definir melhor partição de indivíduos*

   teste = 0
   Para $i = 1,\ldots,n$ fazer

   $k_i = k$ onde $U_{ik} = 1$

   k_f é obtido de acordo com a equação (4.4), assim:

$$k_f = \underset{1\leq l\leq K}{\operatorname{argmin}} \sum_{q=1}^{Q}\sum_{j=1}^{p}(V_{jq})[(a_{ij}-\alpha_{lq})^2+(b_{ij}-\beta_{lq})^2]$$

   se $k_i \neq k_f$, então

   teste = 1

   $U_{ik_f} = 1$

   $U_{ik_i} = 0$

4) *Passo 3: Calcular protótipos*

Computar os componentes $g_{kq} = [\alpha_{kq}, \beta_{kq}]$ dos protótipos $g_{kq}(k = 1, \ldots, K)(q = 1, \ldots, Q)$ de acordo com a equação (4.2) para a nova partição U_{ik} calculada no passo anterior, assim:

$$\alpha_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})a_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})} \quad e \quad \beta_{kj} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})b_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p}(U_{ik})(V_{jq})}$$

5) *Passo 4: Definir melhor partição de variáveis*

teste2 = 0

Para $j = 1, \ldots, p$ fazer

$q_i = q$ onde $V_{jq} = 1$

q_f é obtido de acordo com a equação (4.4), assim:

$$q_f = \underset{1 \le l \le Q}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{n}(U_{ik})[(a_{ij} - \alpha_{kl})^2 + (b_{ij} - \beta_{kl})^2]$$

se $q_i \neq q_f$, então

teste2 = 1

$V_{jq_f} = 1$

$V_{jq_i} = 0$

6) *Critério de parada*

Se teste = 0 e teste2 = 0 então PARE, caso contrário volte ao item (2)

4.2 Análise experimental

O algoritmo de BCKM para variáveis de tipo intervalo será avaliado através de 3 bases de dados reais: Car models, Freshwater fish e City temperatures, tabela 4.1, e sua performance através dos índices OERC e CR será comparada com o modelo KM. Assim, em cada uma das bases o algoritmo é aplicado 1000 vezes, é escolhida a melhor repetição de acordo com o critério de adequação (J) descrito na equação (4.1) e cada base de dados de tipo intervalo é normalizada para valores entre 0 e 1 de acordo com a fomula abaixo, onde a'_{ij} é o valor nomalizado de a_{ij}:

$$a'_{ij} = \frac{a_{ij} - \underset{1 \le l \le n}{\operatorname{argmin}}(a_{lj})}{\underset{1 \le l \le n}{\operatorname{argmax}}(a_{lj}) - \underset{1 \le l \le n}{\operatorname{argmin}}(a_{lj})} \quad e \quad b'_{ij} = \frac{b_{ij} - \underset{1 \le l \le n}{\operatorname{argmin}}(b_{lj})}{\underset{1 \le l \le n}{\operatorname{argmax}}(b_{lj}) - \underset{1 \le l \le n}{\operatorname{argmin}}(b_{lj})} \tag{4.5}$$

Tabela 4.1 Descrição geral das bases de dados reais do tipo intervalo utilizadas

	Car models	Freshwater fish	City temperatures
Qtd. Padrões	33	12	37
Qtd. Variáveis	8	13	12
Qtd. Classes à priori	4	4	-

4.2.1 Car models

Essa base de dados possui 33 modelos de carros que são divididos em 4 categorias de tamanhos diferentes: Utilitarian (10 carros), Berlina (8 carros), Sporting (7 carros) e Luxury (8 carros). Cada carro é descrito por 8 variáveis do tipo intervalo que são: Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width e Height. Podemos ver a descrição da base para as 4 primeiras variáveis na tabela 4.7.Para análise consideramos 8 testes com o mesmo número de clusters de indivíduos, igual a 4, mas se diferenciando na quantidade de clusters de variáveis, assim cada teste, de 1 à 8, descreve respectivamente a quantidade de cluster de variáveis de 1 à 8.

Com os resultados descritos na tabela 4.2 vemos que ao aumentar a quantidade de clusters de variáveis há uma melhor distribuição das variáveis em cada grupo, assim as variáveis com valores similares tendem a ficar no mesmo grupo enquanto as variáveis mais dissimilares vão se separando de acordo com a quantidade de grupos disponíveis. Analisando a performance percebemos pela tabela 4.3 que há uma melhora de acordo com o aumento do número de clusters de variáveis até ocorrer uma estabilização em deteminada quantidade, que no caso desta base de dados foi com a configuração com 4 grupos de variáveis, onde a partir dela temos a mesma performance do algoritmo KM até ser atingida a quantidade de grupos de variáveis igual ao número de variáveis.

Para a configuração com 4 grupos de variáveis podemos ver a divisão dos blocos para os valores de mínimo e máximo nas tabelas 4.8 e 4.9, respectivamente, os critérios de adequação associados a cada bloco na tabela 4.4 e os protótipos obtidos na tabela 4.5.

Tabela 4.2 Descrição da alocação das variáveis nos grupos de variável em cada um dos testes na base de dados Car models

	Quantidade de grupos de variáveis							
	1	2	3	4	5	6	7	8
H_1	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$	x_4, x_8	x_4, x_8	x_8	x_8	x_8	x_8	x_8
H_2	-	$x_1, x_2, x_3, x_5, x_6, x_7$	x_1, x_2, x_3, x_7	x_4	x_4	x_4	x_4	x_4
H_3	-	-	x_5, x_6	x_1, x_2, x_3, x_7	x_1, x_2, x_3, x_7	x_1, x_2	x_1	x_1
H_4	-	-	-	x_5, x_6	x_5	x_5	x_5	x_5
H_5	-	-	-	-	x_6	x_6	x_6	x_6
H_6	-	-	-	-	-	x_3, x_7	x_3, x_7	x_3
H_7	-	-	-	-	-	-	x_2	x_2
H_8	-	-	-	-	-	-	-	x_7

Tabela 4.3 Performance dos algoritmos aplicados ao conjunto de dados Car models para diversas configurações de grupos de variáveis

Algoritmo	KM	BCKM							
Qtd. Grupos Var.	-	1	2	3	4	5	6	7	8
OERC	0.181	0.333	0.303	0.303	0.181	0.181	0.181	0.181	0.181
CR	0.562	0.406	0.395	0.395	0.562	0.562	0.562	0.562	0.562

Tabela 4.4 Matriz do critério de adequação para cada bloco (J_{kq}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Car models com 4 grupos de variáveis

	H_1	H_2	H_3	H_4
P_1	**0.0029**	0.0288	1.2648	0.4132
P_2	0.4610	**0.1424**	2.8611	1.1627
P_3	0.2247	**0.1318**	0.2636	0.3089
P_4	**0.0316**	0.2836	1.5403	1.2382

Tabela 4.5 Protótipos (Min,Max) obtidos para cada um dos blocos para o conjunto de dados Car models com 4 grupos de variáveis

	H_1	H_2	H_3	H_4
P_1	(0.8919, 0.8919)	(0.1947, 0.4097)	(0.3485, 0.5976)	(0.7775, 0.8193)
P_2	(0.4131, 0.4479)	(0.0752, 0.1516)	(0.5865, 0.7007)	(0.4463, 0.4572)
P_3	(0.8547, 0.8547)	(0.5792, 0.8445)	(0.0485, 0.1040)	(0.1591, 0.1605)
P_4	(0.8694, 0.8694)	(0.3785, 0.5496)	(0.1986, 0.2774)	(0.4989, 0.4995)

Tabela 4.6 Descrição da base de dados do tipo intervalo CarBis

	Price	Engine Capacity	...	Width	Height
Alfa 145_U	[27806,33596]	[1370,1910]	...	[171,171]	[143,143]
Alfa 156_B	[41593,62291]	[1598,2492]	...	[175,175]	[142,142]
Alfa 166_L	[64499,88760]	[1970,2959]	...	[182,182]	[142,142]
Aston Martin_S	[260500,460000]	[5935,5935]	...	[183,192]	[124,132]
Audi A3_U	[40230,68838]	[1595,1781]	...	[174,174]	[142,142]
Audi A6_B	[68216,140265]	[1781,4172]	...	[181,181]	[145,145]
Audi A8_L	[123849,171417]	[2771,4172]	...	[188,188]	[144,144]
Bmw serie 3_B	[45407,76392]	[1796,2979]	...	[174,174]	[142,142]
Bmw serie 5_L	[70292,198792]	[2171,4398]	...	[180,180]	[144,144]
Bmw serie 7_L	[104892,276792]	[2793,5397]	...	[186,186]	[143,143]
Ferrari_S	[240292,391692]	[3586,5474]	...	[192,192]	[130,130]
Punto_U	[19229,30885]	[1242,1910]	...	[166,166]	[148,148]
Fiesta_U	[19242,24742]	[1242,1753]	...	[163,163]	[132,132]
Focus_B	[27492,34092]	[1596,1753]	...	[170,170]	[143,143]
Honda NSK_S	[205242,215242]	[2977,3179]	...	[175,175]	[129,129]
Lamborghini_S	[413000,423000]	[5992,5992]	...	[204,204]	[111,111]
Lancia Y_U	[19837,29034]	[1242,1242]	...	[169,169]	[144,144]
Lancia K_L	[58806,81306]	[1998,2959]	...	[183,183]	[146,146]
Maserati GT_S	[155000,159500]	[3217,3217]	...	[182,182]	[131,131]
Mercedes SL_S	[132800,262500]	[2799,5987]	...	[181,181]	[129,129]
Mercedes Classe C_B	[55902,115248]	[1998,3199]	...	[173,173]	[143,143]
Mercedes Classe E_L	[69243,389405]	[1998,5439]	...	[180,180]	[144,144]
Mercedes Classe S_L	[128202,394342]	[3199,5786]	...	[186,186]	[144,144]
Nissan Micra_U	[18492,24192]	[998,1348]	...	[160,160]	[144,144]
Corsa_U	[19212,30612]	[973,1796]	...	[165,165]	[144,144]
Vectra_B	[36492,49092]	[1598,2171]	...	[171,171]	[143,143]
Porsche_S	[147704,246412]	[3387,3600]	...	[177,183]	[130,131]
Twingo_U	[16992,23492]	[1149,1149]	...	[163,163]	[142,142]
Rover 25_U	[21492,33042]	[1119,1994]	...	[169,169]	[142,142]
Rover 75_B	[50490,65399]	[1796,2497]	...	[178,178]	[143,143]
Skoda Fabia_U	[19519,32686]	[1397,1896]	...	[165,165]	[145,145]
Skoda Octavia_B	[27419,48679]	[1585,1896]	...	[173,173]	[143,143]
Passat_L	[39676,63455]	[1595,2496]	...	[470,470]	[175,175]

Tabela 4.7 Descrição da base de dados do tipo intervalo CarBis normalizada

	Price	Engine Capacity	...	Width	Height
Alfa 145_U	[0.02,0.04]	[0.08,0.19]	...	[0.25,0.25]	[0.86,0.86]
Alfa 156_B	[0.06,0.1]	[0.12,0.3]	...	[0.34,0.34]	[0.84,0.84]
Alfa 166_L	[0.11,0.16]	[0.2,0.4]	...	[0.5,0.5]	[0.84,0.84]
Aston Martin_S	[0.55,1]	[0.99,0.99]	...	[0.52,0.73]	[0.35,0.57]
Audi A3_U	[0.05,0.12]	[0.12,0.16]	...	[0.32,0.32]	[0.84,0.84]
Audi A6_B	[0.12,0.28]	[0.16,0.64]	...	[0.48,0.48]	[0.92,0.92]
Audi A8_L	[0.24,0.35]	[0.36,0.64]	...	[0.64,0.64]	[0.89,0.89]
Bmw serie 3_B	[0.06,0.13]	[0.16,0.4]	...	[0.32,0.32]	[0.84,0.84]
Bmw serie 5_L	[0.12,0.41]	[0.24,0.68]	...	[0.45,0.45]	[0.89,0.89]
Bmw serie 7_L	[0.2,0.59]	[0.36,0.88]	...	[0.59,0.59]	[0.86,0.86]
Ferrari_S	[0.5,0.85]	[0.52,0.9]	...	[0.73,0.73]	[0.51,0.51]
Punto_U	[0.01,0.03]	[0.05,0.19]	...	[0.14,0.14]	[1,1]
Fiesta_U	[0.01,0.02]	[0.05,0.16]	...	[0.07,0.07]	[0.57,0.57]
Focus_B	[0.02,0.04]	[0.12,0.16]	...	[0.23,0.23]	[0.86,0.86]
Honda NSK_S	[0.42,0.45]	[0.4,0.44]	...	[0.34,0.34]	[0.49,0.49]
Lamborghini_S	[0.89,0.92]	[1,1]	...	[1,1]	[0,0]
Lancia Y_U	[0.01,0.03]	[0.05,0.05]	...	[0.2,0.2]	[0.89,0.89]
Lancia K_L	[0.09,0.15]	[0.2,0.4]	...	[0.52,0.52]	[0.95,0.95]
Maserati GT_S	[0.31,0.32]	[0.45,0.45]	...	[0.5,0.5]	[0.54,0.54]
Mercedes SL_S	[0.26,0.55]	[0.36,1]	...	[0.48,0.48]	[0.49,0.49]
Mercedes Classe C_B	[0.09,0.22]	[0.2,0.44]	...	[0.3,0.3]	[0.86,0.86]
Mercedes Classe E_L	[0.12,0.84]	[0.2,0.89]	...	[0.45,0.45]	[0.89,0.89]
Mercedes Classe S_L	[0.25,0.85]	[0.44,0.96]	...	[0.59,0.59]	[0.89,0.89]
Nissan Micra_U	[0,0.02]	[0,0.07]	...	[0,0]	[0.89,0.89]
Corsa_U	[0.01,0.03]	[0,0.16]	...	[0.11,0.11]	[0.89,0.89]
Vectra_B	[0.04,0.07]	[0.12,0.24]	...	[0.25,0.25]	[0.86,0.86]
Porsche_S	[0.3,0.52]	[0.48,0.52]	...	[0.39,0.52]	[0.51,0.54]
Twingo_U	[0,0.01]	[0.04,0.04]	...	[0.07,0.07]	[0.84,0.84]
Rover 25_U	[0.01,0.04]	[0.03,0.2]	...	[0.2,0.2]	[0.84,0.84]
Rover 75_B	[0.08,0.11]	[0.16,0.3]	...	[0.41,0.41]	[0.86,0.86]
Skoda Fabia_U	[0.01,0.04]	[0.08,0.18]	...	[0.11,0.11]	[0.92,0.92]
Skoda Octavia_B	[0.02,0.07]	[0.12,0.18]	...	[0.3,0.3]	[0.86,0.86]
Passat_L	[0.05,0.1]	[0.12,0.3]	...	[0.34,0.34]	[0.95,0.95]

Tabela 4.8 Matriz de blocos para os valores de mínimo do conjunto de dados Car models com 4 grupos de variáveis

		H_1	H_2	H_3				H_4	
		x_8	x_4	x_1	x_2	x_3	x_7	x_5	x_6
	e_6	0.9189	0.2137	0.1156	0.1610	0.3568	0.4773	0.5541	0.7919
	e_7	0.8919	0.1145	0.2412	0.3582	0.4432	0.6364	0.7297	0.9249
	e_9	0.8919	0.2137	0.1203	0.2387	0.4108	0.4545	0.6487	0.7804
P_1	e_{10}	0.8649	0.2366	0.1984	0.3626	0.4216	0.5909	0.7838	0.8960
	e_{22}	0.8919	0.1374	0.1180	0.2042	0.3892	0.4545	0.6487	0.8035
	e_{23}	0.8919	0.2519	0.2510	0.4435	0.3243	0.5909	0.8378	0.9306
	e_4	0.3514	0.0611	0.5497	0.9886	0.8000	0.5227	0.3243	0.7052
	e_{11}	0.5135	0.0458	0.5041	0.5206	0.7838	0.7273	0.3378	0.7688
	e_{15}	0.4865	0.1374	0.4249	0.3993	0.5946	0.3409	0.2432	0.4104
P_2	e_{16}	0.0000	0.0000	0.8939	1.0000	1.0000	1.0000	0.4054	0.6012
	e_{19}	0.5405	0.0916	0.3115	0.4471	0.7027	0.5000	0.4189	0.6243
	e_{20}	0.4865	0.1679	0.2614	0.3638	0.4432	0.4773	0.2297	0.6012
	e_{27}	0.5135	0.0229	0.2951	0.4810	0.7027	0.3864	0.0000	0.5780
	e_{12}	1.0000	0.6336	0.0051	0.0536	0.0270	0.1364	0.1487	0.2139
	e_{13}	0.5676	0.7023	0.0051	0.0536	0.0919	0.0682	0.1351	0.2312
	e_{17}	0.8919	0.5573	0.0064	0.0536	0.0432	0.2046	0.0405	0.1676
P_3	e_{24}	0.8919	0.6565	0.0034	0.0050	0.0000	0.0000	0.0135	0.1850
	e_{25}	0.8919	0.3893	0.0050	0.0000	0.0270	0.1136	0.1892	0.2254
	e_{28}	0.8378	0.5954	0.0000	0.0351	0.0054	0.0682	0.0000	0.0000
	e_{29}	0.8378	0.5191	0.0102	0.0291	0.0541	0.2046	0.2162	0.3237
	e_{31}	0.9189	0.5802	0.0057	0.0845	0.0378	0.1136	0.1487	0.3064
	e_1	0.8649	0.3359	0.0244	0.0791	0.1892	0.2500	0.2568	0.3642
	e_2	0.8378	0.3511	0.0555	0.1245	0.2703	0.3409	0.3378	0.5780
	e_3	0.8378	0.4504	0.1072	0.1986	0.2919	0.5000	0.4730	0.7457
	e_5	0.8378	0.2214	0.0525	0.1239	0.2108	0.3182	0.2027	0.4162
	e_8	0.8378	0.2061	0.0641	0.1640	0.2757	0.3182	0.5135	0.6012
P_4	e_{14}	0.8649	0.5267	0.0237	0.1241	0.1892	0.2273	0.3649	0.4162
	e_{18}	0.9460	0.3817	0.0944	0.2042	0.3351	0.5227	0.4730	0.7283
	e_{21}	0.8649	0.0992	0.0878	0.2042	0.3243	0.2955	0.5000	0.6358
	e_{26}	0.8649	0.5038	0.0440	0.1245	0.2324	0.2500	0.3919	0.6185
	e_{30}	0.8649	0.4809	0.0756	0.1640	0.2432	0.4091	0.5405	0.7630
	e_{32}	0.8649	0.5496	0.0235	0.1219	0.2162	0.2955	0.2162	0.6301
	e_{33}	0.9460	0.4351	0.0512	0.1239	0.2270	0.3409	0.4730	0.7341

Tabela 4.9 Matriz de blocos para os valores de máximo do conjunto de dados Car models com 4 grupos de variáveis

		H_1	H_2	H_3				H_4	
		x_8	x_4	x_1	x_2	x_3	x_7	x_5	x_6
	e_6	0.9189	0.4428	0.2783	0.6374	0.5405	0.4773	0.5541	0.7919
	e_7	0.8919	0.4733	0.3486	0.6374	0.5405	0.6364	0.7297	0.9249
	e_9	0.8919	0.3970	0.4104	0.6824	0.5405	0.4545	0.6487	0.7804
P_1	e_{10}	0.8649	0.3588	0.5864	0.8815	0.4865	0.5909	0.9730	0.9769
	e_{22}	0.8919	0.4428	0.8407	0.8898	0.5405	0.4545	0.6487	0.8035
	e_{23}	0.8919	0.3435	0.8518	0.9590	0.4865	0.5909	1.0000	1.0000
	e_4	0.5676	0.0840	1.0000	0.9886	0.8432	0.7273	0.4595	0.7168
	e_{11}	0.5135	0.0992	0.8458	0.8968	0.8000	0.7273	0.3378	0.7688
	e_{15}	0.4865	0.1985	0.4475	0.4395	0.6487	0.3409	0.2432	0.4104
P_2	e_{16}	0.0000	0.0000	0.9165	1.0000	1.0000	1.0000	0.4054	0.6012
	e_{19}	0.5405	0.1374	0.3217	0.4471	0.7568	0.5000	0.4189	0.6243
	e_{20}	0.4865	0.4428	0.5542	0.9990	0.5405	0.4773	0.2297	0.6012
	e_{27}	0.5405	0.0992	0.5179	0.5234	0.8378	0.5227	0.0000	0.5838
	e_{12}	1.0000	0.7939	0.0314	0.1867	0.1081	0.1364	0.1487	0.2370
	e_{13}	0.5676	0.7634	0.0175	0.1554	0.0919	0.0682	0.1351	0.2312
	e_{17}	0.8919	0.7786	0.0272	0.0536	0.1297	0.2046	0.0405	0.1676
P_3	e_{24}	0.8919	0.8855	0.0163	0.0747	0.0757	0.0000	0.0135	0.1850
	e_{25}	0.8919	1.0000	0.0307	0.1640	0.2811	0.1136	0.1892	0.2254
	e_{28}	0.8378	0.7252	0.0147	0.0351	0.0973	0.0682	0.0000	0.0000
	e_{29}	0.8378	0.8473	0.0362	0.2034	0.1892	0.2046	0.2162	0.3237
	e_{31}	0.9189	0.9618	0.0354	0.1839	0.1784	0.1136	0.1487	0.3064
	e_1	0.8649	0.5573	0.0375	0.1867	0.3297	0.2500	0.2568	0.3642
	e_2	0.8378	0.5038	0.1023	0.3027	0.4162	0.3409	0.3378	0.5780
	e_3	0.8378	0.4580	0.1620	0.3957	0.3297	0.5000	0.4730	0.7457
	e_5	0.8378	0.5344	0.1170	0.1610	0.4757	0.3182	0.2162	0.4162
	e_8	0.8378	0.5344	0.1341	0.3997	0.5243	0.3182	0.5135	0.6012
P_4	e_{14}	0.8649	0.5420	0.0386	0.1554	0.2324	0.2273	0.3649	0.4162
	e_{18}	0.9460	0.4046	0.1452	0.3957	0.3784	0.5227	0.4730	0.7283
	e_{21}	0.8649	0.5420	0.2218	0.4435	0.5405	0.2955	0.5000	0.6358
	e_{26}	0.8649	0.6565	0.0725	0.2387	0.3081	0.2500	0.3919	0.6185
	e_{30}	0.8649	0.5878	0.1093	0.3037	0.3243	0.4091	0.5405	0.7630
	e_{32}	0.8649	0.6031	0.0715	0.1839	0.2216	0.2955	0.2162	0.6301
	e_{33}	0.9460	0.6718	0.1049	0.3035	0.3784	0.3409	0.4730	0.7341

4.2.2 Freshwater fish

Essa base de dados possui 12 espécies de peixes de água doce descritas pelas 13 seguintes variáveis: Length, Weight, Muscle, Intestine, Stomach, Gills, Liver, Kidneys, Liver/Muscle, Kidneys/Muscle, Gills/Muscle, Intestine/Muscle e Stomach/Muscle. Ela possui 4 classes à

priori que dividem desigualmente as 12 instâncias de acordo com a dieta: Carnivorous (4 instâncias), Detritivorous (4 instâncias), Omnivorous (2 instâncias) e Herbivorous (2 instâncias). A descrição da base Freshwater fish pode ser vista na tabela 4.13 e sua equivalente normalizada na tabela 4.14. Foram feitos testes para configurações com 2, 4, 6 e 9 grupos de variáveis, onde todos possuem 4 grupos de indivíduos.

Os resultados obtidos na tabela 4.10 mostram que o algoritmo BCKM para uma configuração com 2 clusters de variáveis traz uma performance melhor que o KM, consequentemente os indivíduos são agrupados mais precisamente. Pela tabela 4.11 notamos que as variáveis vão se adaptando a quantidade de clusters disponíveis, se agrupando com as variáveis que sejam mais similares e ficando em grupos diferentes das variáveis mais dissimilares a ela. Assim, para a configuração com 2 clusters de variáveis, podemos constatar a maior similaridade entre as variáveis de indivíduos no mesmo bloco quando comparado aos que estão em blocos diferentes pelas tabelas 4.15 e 4.16, que trazem os valores de mínimo e máximo, respectivamente, das variáveis de tipo intervalo. Os índices de adequação para cada bloco estão na tabela 4.12, trazendo destaque em negrito para o grupo de variáveis mais homogêneo em cada grupo de indivíduos e sublinhado para o grupo de indivíduos mais homogêneo em cada grupo de variáveis. Os protótipos que representam cada bloco estão na tabela 4.17.

Tabela 4.10 Performance dos algoritmos aplicados ao conjunto de dados Freshwater fish para diversos testes

Algoritmo	KM	BCKM			
Qtd. Grupos Var.	-	2	4	6	9
OERC	0.416	0.250	0.333	0.416	0.416
CR	0.067	0.313	0.237	0.067	0.067

Tabela 4.11 Descrição da alocação das variáveis nos grupos de variável em cada uma das configurações de grupos de variáveis testadas para base de dados Freshwater fish

	Quantidade de grupos de variáveis			
	2	4	6	9
H_1	x_9, x_{10}, x_{11}, x_{12}, x_{13}	x_9,x_{10}, x_{12}	x_{10},x_{12}	x_{10},x_{12}
H_2	x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8	x_3, x_5, x_6, x_7, x_8	x_5, x_6, x_7	x_5, x_6, x_7
H_3	-	x_1, x_2, x_4	x_1, x_2	x_1
H_4	-	x_{11}, x_{13}	x_{11}, x_{13}	x_{11}, x_{13}
H_5	-	-	x_3, x_4, x_8	x_3
H_6	-	-	x_9	x_4
H_7	-	-	-	x_8
H_8	-	-	-	x_9
H_9	-	-	-	x_2

Tabela 4.12 Matriz do critério de adequação para cada bloco (J_{kq}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados Freshwater fish com 2 grupos de variáveis

	H_1	H_2
P_1	**0.9647**	2.7808
P_2	2.1230	**0.5485**
P_3	**0.0542**	1.9925
P_4	0.3343	**0.0099**

Tabela 4.13 Descrição da base de dados do tipo intervalo Freshwater fish

	Length	Weight	...	Intestine/Muscle	Stomach/Muscle
Ageneiosusbrevifili	[22.5,35.5]	[170,625]	...	[0.23,0.63]	[0,0.55]
Cynodongibbus	[19,32]	[77,359]	...	[0,0.5]	[0.2,1.24]
Hopliasaïmara	[25.5,63]	[340,5500]	...	[0.11,0.49]	[0.09,0.4]
Potamotrygonhystrix	[20.5,45]	[400,6250]	...	[0,1.25]	[0,0.5]
Leporinusfasciatus	[18.8,25]	[125,273]	...	[0,0]	[0.12,0.17]
Leporinusfrederici	[23,24.5]	[290,350]	...	[0.18,0.24]	[0.13,0.58]
Dorasmicropoeus	[19.2,31]	[128,505]	...	[0,1.48]	[0,0.79]
Platydorascostatus	[13.7,25]	[60,413]	...	[0.3,1.45]	[0,0.61]
Pseudoancistrusbarbatus	[13,20.5]	[55,210]	...	[0,2.31]	[0.49,1.36]
Semaprochilodusvari	[22,28]	[330,700]	...	[0.4,1.68]	[0,1.25]
Acnodonoligacanthus	[10,16.2]	[34.9,154.7]	...	[0,2.16]	[0.23,5.97]
Myleusrubripinis	[12.3,18]	[80,275]	...	[0,0]	[0.31,4.33]

Tabela 4.14 Descrição da base de dados do tipo intervalo Freshwater fish normalizada

	Length	Weight	...	Intestine/Muscle	Stomach/Muscle
Ageneiosusbrevifili	[0.24,0.48]	[0.02,0.09]	...	[0.1,0.27]	[0,0.09]
Cynodongibbus	[0.17,0.42]	[0.01,0.05]	...	[0,0.22]	[0.03,0.21]
Hopliasaïmara	[0.29,1]	[0.05,0.88]	...	[0.05,0.21]	[0.02,0.07]
Potamotrygonhystrix	[0.2,0.66]	[0.06,1]	...	[0,0.54]	[0,0.08]
Leporinusfasciatus	[0.17,0.28]	[0.01,0.04]	...	[0,0]	[0.02,0.03]
Leporinusfrederici	[0.25,0.27]	[0.04,0.05]	...	[0.08,0.1]	[0.02,0.1]
Dorasmicropoeus	[0.17,0.4]	[0.01,0.08]	...	[0,0.64]	[0,0.13]
Platydorascostatus	[0.07,0.28]	[0,0.06]	...	[0.13,0.63]	[0,0.1]
Pseudoancistrusbarbatus	[0.06,0.2]	[0,0.03]	...	[0,1]	[0.08,0.23]
Semaprochilodusvari	[0.23,0.34]	[0.05,0.11]	...	[0.17,0.73]	[0,0.21]
Acnodonoligacanthus	[0,0.12]	[0,0.02]	...	[0,0.94]	[0.04,1]
Myleusrubripinis	[0.04,0.15]	[0.01,0.04]	...	[0,0]	[0.05,0.73]

Tabela 4.15 Matriz de blocos para os valores de mínimo do conjunto de dados Freshwater fish com 2 grupos de variáveis

		H_1					H_2							
		x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	e_1	0.00	0.00	0.04	0.10	0.00	0.24	0.02	0.16	0.11	0.00	0.20	0.01	0.00
	e_4	0.03	0.10	0.00	0.00	0.00	0.20	0.06	0.08	0.00	0.00	0.00	0.03	0.05
P_1	e_5	0.00	0.03	0.00	0.00	0.02	0.17	0.01	0.14	0.00	0.02	0.00	0.01	0.04
	e_6	0.00	0.02	0.01	0.08	0.02	0.25	0.04	0.03	0.02	0.00	0.01	0.00	0.00
	e_7	0.04	0.00	0.03	0.00	0.00	0.17	0.01	0.09	0.00	0.00	0.04	0.05	0.00
	e_8	0.02	0.04	0.01	0.13	0.00	0.07	0.00	0.04	0.07	0.00	0.02	0.03	0.03
	e_9	0.06	0.00	0.00	0.00	0.08	0.06	0.00	0.01	0.00	0.01	0.00	0.01	0.00
P_2	e_{10}	0.28	0.14	0.05	0.17	0.00	0.23	0.05	0.03	0.05	0.00	0.03	0.12	0.03
	e_{12}	0.13	0.11	0.00	0.00	0.05	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	e_2	0.01	0.03	0.04	0.00	0.03	0.17	0.01	0.27	0.00	0.04	0.18	0.06	0.05
P_3	e_3	0.01	0.04	0.04	0.05	0.02	0.29	0.05	0.14	0.15	0.04	0.14	0.02	0.04
P_4	e_{11}	0.02	0.08	0.02	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Tabela 4.16 Matriz de blocos para os valores de máximo do conjunto de dados Freshwater fish com 2 grupos de variáveis

		H_1					H_2							
		x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	e_1	0.02	0.09	0.09	0.27	0.09	0.48	0.09	0.58	1.00	0.16	0.43	0.16	0.14
	e_4	0.14	0.39	0.00	0.54	0.08	0.66	1.00	0.11	0.31	0.04	0.00	0.16	0.29
P_1	e_5	0.00	0.06	0.02	0.00	0.03	0.28	0.04	0.23	0.00	0.02	0.07	0.02	0.10
	e_6	0.01	0.04	0.02	0.10	0.10	0.27	0.05	0.12	0.09	0.06	0.02	0.01	0.03
	e_7	0.15	0.67	0.05	0.64	0.13	0.40	0.08	0.22	0.48	0.12	0.10	0.24	0.71
	e_8	0.21	0.72	0.08	0.63	0.10	0.28	0.06	0.10	0.22	0.03	0.09	0.14	0.27
	e_9	0.32	0.02	0.00	1.00	0.23	0.20	0.03	0.01	0.07	0.01	0.00	0.04	0.00
P_2	e_{10}	1.00	0.31	0.07	0.73	0.21	0.34	0.11	0.06	0.31	0.06	0.05	0.52	0.10
	e_{12}	0.57	0.47	0.25	0.00	0.73	0.15	0.04	0.00	0.00	0.00	0.00	0.01	0.00
	e_2	0.09	0.15	0.07	0.22	0.21	0.42	0.05	1.00	0.89	1.00	1.00	1.00	1.00
P_3	e_3	0.04	0.23	0.07	0.21	0.07	1.00	0.88	0.68	0.73	0.10	0.42	0.14	0.80
P_4	e_{11}	0.34	1.00	1.00	0.94	1.00	0.12	0.02	0.01	0.03	0.02	0.07	0.02	0.02

Tabela 4.17 Protótipos (Min,Max) obtidos para cada um dos blocos para o conjunto de dados Freshwater fish com 2 grupos de variáveis

	H_1	H_2
P_1	(0.0208, 0.1505)	(0.0565, 0.2352)
P_2	(0.0643, 0.3824)	(0.0289, 0.1040)
P_3	(0.0268, 0.1343)	(0.1036, 0.6945)
P_4	(0.0306, 0.8558)	(0.0014, 0.0386)

4.2.3 City temperatures

A base de dados do tipo intervalo City temperatures possui a descrição mínima e máxima das temperaturas de 37 cidades ao redor do mundo para cada mês do ano, então temos 12 variáveis. Assim, foram feitos testes para cada uma das configurações possíveis de agrupamento de variáveis, de 1 a 12 grupos de variáveis, mantendo 4 grupos de indivíduos para todos os testes.

A partição de indivíduos obtida pelo algoritmo KM nessa base com 4 clusters de indivíduos é igual a gerada, ao final da execução, pelo algoritmo BCKM, para qualquer configuração de clusters de variáveis (de 1 à 12), assim este algoritmo não altera o comportamento da base a nível de clusters de indivíduos e podemos focar na descrição dos grupos de variáveis.

Percebemos pela figura 4.1 que a partir da divisão da partição de variáveis em 3 clusters a variável 5 (maio) fica isolada mostrando a sua dissimilaridade em relação as outras variáveis, que realmente é confirmada quando vemos a descrição da base, seja ela normalizada, tabela 4.21, ou sem normalização, tabela 4.20. Apesar disso, em geral, a divisão das variáveis nos grupos para a configuração de 4 clusters fica de acordo com as estações do ano como podemos ver na figura 4.1. Nela o mês de Maio fica sozinha em um grupo, dada sua dissimilaridade, os meses de dezembro a março se incluem em outro grupo, que para os países do hemisfério norte seria inverno e para os do sul seria verão, onde em cada bloco, de acordo com o número de grupos de indivíduos, é melhor visualizada a proximidade das temperaturas. Um terceiro grupo de variáveis tem os meses de abril, outubro e novembro, que é aceitável já que esses meses são de transição de temperaturas quentes para temperaturas frias ou de temperaturas frias para temperaturas quentes que virão nos meses seguintes. O último grupo de variáveis formado inclui os meses de junho a setembro que, também, representam o verão ou o inverno dependendo da localização da cidade em relação a linha do equador.

Os índividuos e variáveis inclusos em cada bloco ficam agrupados, tanto por sua proximidade a nível geral de temperaturas ao longo do ano como por suas similaridade em temperaturas em determinados meses, as quais podemos ver que são mais homogêneas em determinados blocos, como mostra a tabela 4.22 para os valores de mínimo e a tabela 4.22 para os valores de máximo da divisão dos blocos de acordo com o índice de adequação para a configuração com 4 clusters de variáveis na tabela 4.18, que destaca em negrito o grupo de variáveis mais homogêneo em cada grupo de indivíduos e é sublinhado o grupo de indivíduos mais homogêneo para cada um dos grupos de variáveis.

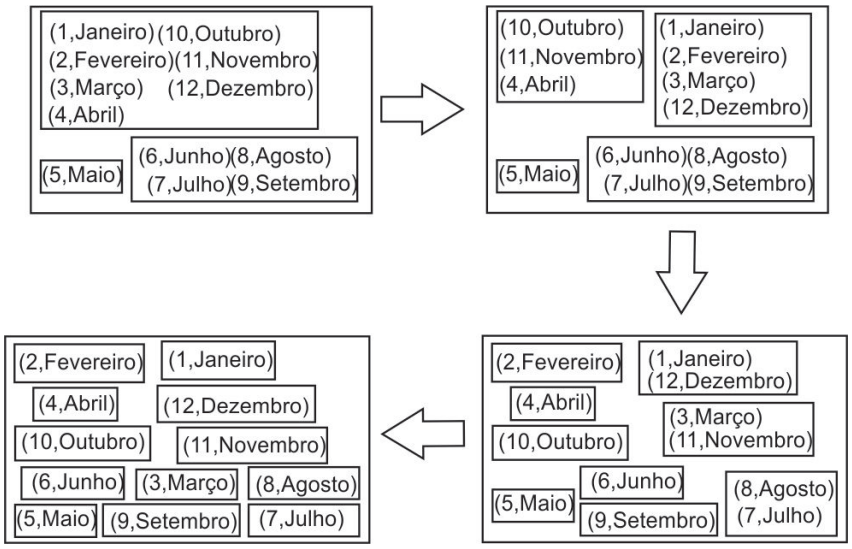

Figura 4.1 Progresso da alocação das variáveis nas configurações com 3, 4, 9 e 12 grupos de variável para base de dados City temperatures

Tabela 4.18 Matriz do critério de adequação para cada bloco (J_{kq}) obtida pela aplicação do algoritmo de partição BCKM ao conjunto de dados City temperatures com 4 grupos de variáveis

	H_1	H_2	H_3	H_4
P_1	**0.0286**	0.2818	0.1287	0.5334
P_2	**0.1887**	1.2803	0.7803	0.8253
P_3	**0.0501**	0.5197	0.3560	1.8107
P_4	**0.1056**	1.1465	0.4869	0.6808

Tabela 4.19 Protótipos (Min,Max) obtidos para cada um dos blocos para o conjunto de dados City temperaturess com 4 grupos de variáveis

	H_1	H_2	H_3	H_4
P_1	(0.4896, 0.6562)	(0.6430, 0.8705)	(0.4983, 0.7490)	(0.2222, 0.4897)
P_2	(0.2847, 0.5625)	(0.1789, 0.3754)	(0.1410, 0.4092)	(0.1496, 0.4988)
P_3	(0.4421, 0.6389)	(0.3460, 0.5163)	(0.3311, 0.5687)	(0.3795, 0.6737)
P_4	(0.6701, 0.8576)	(0.6653, 0.8744)	(0.6511, 0.8974)	(0.5771, 0.8090)

Tabela 4.20 Descrição da base de dados do tipo intervalo City temperatures

	Janeiro	Fevereiro	...	Maio	...	Novembro	Dezembro
Amssterdam	[-4,4]	[4,-5]	...	[7,17]	...	[1,10]	[-1,4]
Athens	[6,12]	[12,6]	...	[16,25]	...	[11,18]	[8,14]
Bahrain	[13,19]	[19,14]	...	[25,32]	...	[20,26]	[15,21]
Bombay	[19,28]	[28,19]	...	[27,33]	...	[23,32]	[20,30]
Cairo	[8,20]	[20,9]	...	[17,33]	...	[14,26]	[10,20]
Calcutta	[13,27]	[27,16]	...	[26,36]	...	[18,29]	[13,26]
Colombo	[22,30]	[30,22]	...	[25,31]	...	[23,29]	[22,30]
Copenhagen	[-2,2]	[2,-3]	...	[8,16]	...	[3,7]	[1,4]
Dubal	[13,23]	[23,14]	...	[22,34]	...	[17,30]	[14,26]
Frankfurt	[-10,9]	[9,-8]	...	[3,27]	...	[-3,14]	[-8,10]
Geneva	[-3,5]	[5,-6]	...	[10,17]	...	[3,8]	[-2,6]
HongKong	[13,17]	[17,12]	...	[22,27]	...	[18,23]	[14,19]
KulaLumpur	[22,31]	[31,23]	...	[23,32]	...	[23,31]	[23,31]
Lisbon	[8,13]	[13,8]	...	[13,21]	...	[11,17]	[8,14]
London	[2,6]	[6,2]	...	[8,17]	...	[5,10]	[3,7]
Madras	[20,30]	[30,20]	...	[28,39]	...	[22,30]	[21,29]
Madrid	[1,9]	[9,1]	...	[9,24]	...	[4,14]	[1,9]
Manila	[21,27]	[27,22]	...	[25,31]	...	[22,28]	[22,27]
Mauritius	[22,28]	[28,22]	...	[19,25]	...	[19,27]	[21,28]
MexicoCity	[6,22]	[22,15]	...	[18,27]	...	[14,25]	[8,23]
Moscow	[-13,-6]	[-6,-15]	...	[7,18]	...	[-5,0]	[-11,-5]
Munich	[-6,1]	[1,-5]	...	[7,18]	...	[0,7]	[-4,2]
Nairobi	[12,25]	[25,13]	...	[13,22]	...	[12,23]	[13,23]
NewDelhi	[6,21]	[21,10]	...	[26,40]	...	[11,28]	[7,23]
NewYork	[-2,4]	[4,-3]	...	[12,22]	...	[5,12]	[-2,6]
Paris	[1,7]	[7,1]	...	[8,19]	...	[4,10]	[1,6]
Rome	[4,11]	[11,5]	...	[13,23]	...	[9,16]	[5,12]
SanFrancisco	[6,13]	[13,6]	...	[10,19]	...	[8,18]	[6,14]
Seoul	[0,7]	[7,1]	...	[12,22]	...	[7,19]	[1,8]
Singapore	[23,30]	[30,23]	...	[24,30]	...	[24,30]	[23,30]
Stockholm	[-9,-5]	[-5,-9]	...	[6,15]	...	[1,4]	[-2,2]
Sydney	[20,30]	[30,20]	...	[12,20]	...	[16,26]	[20,30]
Tehran	[0,5]	[5,5]	...	[20,25]	...	[9,12]	[-5,0]
Tokyo	[0,9]	[9,0]	...	[14,23]	...	[8,16]	[2,12]
Toronto	[-8,-1]	[-1,-8]	...	[-8,18]	...	[-1,17]	[-5,1]
Vienna	[-2,1]	[1,-1]	...	[10,19]	...	[2,7]	[1,3]
Zurich	[-11,9]	[9,-8]	...	[2,27]	...	[0,19]	[-11,8]

Tabela 4.21 Descrição da base de dados do tipo intervalo City temperatures normalizada

	Janeiro	Fevereiro	...	Maio	...	Novembro	Dezembro
Amssterdam	[0.2,0.39]	[0.21,0.38]	...	[0.31,0.52]	...	[0.16,0.41]	[0.24,0.36]
Athens	[0.43,0.57]	[0.45,0.57]	...	[0.5,0.69]	...	[0.43,0.62]	[0.45,0.6]
Bahrain	[0.59,0.73]	[0.62,0.72]	...	[0.69,0.83]	...	[0.68,0.84]	[0.62,0.76]
Bombay	[0.73,0.93]	[0.72,0.91]	...	[0.73,0.85]	...	[0.76,1]	[0.74,0.98]
Cairo	[0.48,0.75]	[0.51,0.79]	...	[0.52,0.85]	...	[0.51,0.84]	[0.5,0.74]
Calcutta	[0.59,0.91]	[0.66,0.94]	...	[0.71,0.92]	...	[0.62,0.92]	[0.57,0.88]
Colombo	[0.8,0.98]	[0.79,0.96]	...	[0.69,0.81]	...	[0.76,0.92]	[0.79,0.98]
Copenhagen	[0.25,0.34]	[0.26,0.36]	...	[0.33,0.5]	...	[0.22,0.32]	[0.29,0.36]
Dubal	[0.59,0.82]	[0.62,0.83]	...	[0.63,0.88]	...	[0.59,0.95]	[0.6,0.88]
Frankfurt	[0.07,0.5]	[0.15,0.66]	...	[0.23,0.73]	...	[0.05,0.51]	[0.07,0.5]
Geneva	[0.23,0.41]	[0.19,0.45]	...	[0.38,0.52]	...	[0.22,0.35]	[0.21,0.4]
HongKong	[0.59,0.68]	[0.57,0.66]	...	[0.63,0.73]	...	[0.62,0.76]	[0.6,0.71]
KulaLumpur	[0.8,1]	[0.81,1]	...	[0.65,0.83]	...	[0.76,0.97]	[0.81,1]
Lisbon	[0.48,0.59]	[0.49,0.62]	...	[0.44,0.6]	...	[0.43,0.59]	[0.45,0.6]
London	[0.34,0.43]	[0.36,0.47]	...	[0.33,0.52]	...	[0.27,0.41]	[0.33,0.43]
Madras	[0.75,0.98]	[0.74,0.98]	...	[0.75,0.98]	...	[0.73,0.95]	[0.76,0.95]
Madrid	[0.32,0.5]	[0.34,0.57]	...	[0.35,0.67]	...	[0.24,0.51]	[0.29,0.48]
Manila	[0.77,0.91]	[0.79,0.89]	...	[0.69,0.81]	...	[0.73,0.89]	[0.79,0.9]
Mauritius	[0.8,0.93]	[0.79,0.94]	...	[0.56,0.69]	...	[0.65,0.86]	[0.76,0.93]
MexicoCity	[0.43,0.8]	[0.64,0.81]	...	[0.54,0.73]	...	[0.51,0.81]	[0.45,0.81]
Moscow	[0,0.16]	[0,0.06]	...	[0.31,0.54]	...	[0,0.14]	[0,0.14]
Munich	[0.16,0.32]	[0.21,0.38]	...	[0.31,0.54]	...	[0.14,0.32]	[0.17,0.31]
Nairobi	[0.57,0.86]	[0.6,0.87]	...	[0.44,0.63]	...	[0.46,0.76]	[0.57,0.81]
NewDelhi	[0.43,0.77]	[0.53,0.83]	...	[0.71,1]	...	[0.43,0.89]	[0.43,0.81]
NewYork	[0.25,0.39]	[0.26,0.4]	...	[0.42,0.63]	...	[0.27,0.46]	[0.21,0.4]
Paris	[0.32,0.45]	[0.34,0.47]	...	[0.33,0.56]	...	[0.24,0.41]	[0.29,0.4]
Rome	[0.39,0.55]	[0.43,0.6]	...	[0.44,0.65]	...	[0.38,0.57]	[0.38,0.55]
SanFrancisco	[0.43,0.59]	[0.45,0.62]	...	[0.38,0.56]	...	[0.35,0.62]	[0.4,0.6]
Seoul	[0.3,0.45]	[0.34,0.45]	...	[0.42,0.63]	...	[0.32,0.65]	[0.29,0.45]
Singapore	[0.82,0.98]	[0.81,0.96]	...	[0.67,0.79]	...	[0.78,0.95]	[0.81,0.98]
Stockholm	[0.09,0.18]	[0.13,0.19]	...	[0.29,0.48]	...	[0.16,0.24]	[0.21,0.31]
Sydney	[0.75,0.98]	[0.74,0.96]	...	[0.42,0.58]	...	[0.57,0.84]	[0.74,0.98]
Tehran	[0.3,0.41]	[0.43,0.49]	...	[0.58,0.69]	...	[0.38,0.46]	[0.14,0.26]
Tokyo	[0.3,0.5]	[0.32,0.53]	...	[0.46,0.65]	...	[0.35,0.57]	[0.31,0.55]
Toronto	[0.11,0.27]	[0.15,0.3]	...	[0,0.54]	...	[0.11,0.59]	[0.14,0.29]
Vienna	[0.25,0.32]	[0.3,0.38]	...	[0.38,0.56]	...	[0.19,0.32]	[0.29,0.33]
Zurich	[0.05,0.5]	[0.15,0.64]	...	[0.21,0.73]	...	[0.14,0.65]	[0,0.45]

Tabela 4.22 Matriz de blocos para os valores de mínimo do conjunto de dados City temperatures normalizado com 4 grupos de variáveis

		H_1	H_2				H_3			H_4			
		Mai.	Jan.	Fev.	Mar.	Dez.	Abr.	Out.	Nov.	Jun.	Jul.	Ago.	Set.
	e_{19}	0.56	0.80	0.79	0.71	0.76	0.61	0.53	0.65	0.38	0.29	0.28	0.38
	e_{20}	0.54	0.43	0.64	0.60	0.45	0.53	0.47	0.51	0.38	0.32	0.31	0.41
P_1	e_{23}	0.44	0.57	0.60	0.52	0.57	0.42	0.38	0.46	0.21	0.10	0.09	0.19
	e_{32}	0.42	0.75	0.74	0.62	0.74	0.47	0.38	0.57	0.00	0.00	0.03	0.19
	e_1	0.31	0.20	0.21	0.24	0.24	0.18	0.15	0.16	0.15	0.06	0.13	0.16
	e_8	0.33	0.25	0.26	0.17	0.29	0.13	0.21	0.22	0.18	0.19	0.19	0.19
	e_{10}	0.23	0.07	0.15	0.10	0.07	0.05	0.00	0.05	0.06	0.00	0.00	0.00
	e_{11}	0.38	0.23	0.19	0.26	0.21	0.24	0.18	0.22	0.29	0.26	0.25	0.19
	e_{15}	0.33	0.34	0.36	0.26	0.33	0.18	0.24	0.27	0.18	0.16	0.16	0.19
P_2	e_{21}	0.31	0.00	0.00	0.00	0.00	0.05	0.03	0.00	0.18	0.16	0.09	0.03
	e_{22}	0.31	0.16	0.21	0.14	0.17	0.13	0.12	0.14	0.15	0.13	0.09	0.09
	e_{26}	0.33	0.32	0.34	0.24	0.29	0.18	0.21	0.24	0.21	0.19	0.16	0.19
	e_{31}	0.29	0.09	0.13	0.10	0.21	0.08	0.15	0.16	0.18	0.19	0.16	0.13
	e_{35}	0.00	0.11	0.15	0.10	0.14	0.00	0.18	0.11	0.24	0.26	0.25	0.22
	e_{36}	0.38	0.25	0.30	0.21	0.29	0.18	0.21	0.19	0.24	0.23	0.19	0.19
	e_{37}	0.21	0.05	0.15	0.02	0.00	0.03	0.09	0.14	0.03	0.06	0.00	0.00
	e_2	0.50	0.43	0.45	0.38	0.45	0.34	0.47	0.43	0.41	0.45	0.44	0.44
	e_{14}	0.44	0.48	0.49	0.38	0.45	0.34	0.41	0.43	0.32	0.29	0.31	0.38
	e_{17}	0.35	0.32	0.34	0.26	0.29	0.21	0.24	0.24	0.24	0.26	0.25	0.25
	e_{25}	0.42	0.25	0.26	0.21	0.21	0.21	0.32	0.27	0.35	0.42	0.31	0.34
P_3	e_{27}	0.44	0.39	0.43	0.36	0.38	0.32	0.38	0.38	0.35	0.39	0.38	0.38
	e_{28}	0.38	0.43	0.45	0.36	0.40	0.26	0.32	0.35	0.18	0.13	0.13	0.22
	e_{29}	0.42	0.30	0.34	0.21	0.29	0.21	0.09	0.32	0.32	0.32	0.25	0.13
	e_{33}	0.58	0.30	0.43	0.43	0.14	0.45	0.53	0.38	0.68	0.90	0.94	0.75
	e_{34}	0.46	0.30	0.32	0.26	0.31	0.29	0.38	0.35	0.38	0.45	0.47	0.47
	e_3	0.69	0.59	0.62	0.60	0.62	0.61	0.71	0.68	0.68	0.68	0.69	0.72
	e_4	0.73	0.73	0.72	0.71	0.74	0.68	0.71	0.76	0.62	0.55	0.53	0.59
	e_5	0.52	0.48	0.51	0.45	0.50	0.42	0.53	0.51	0.44	0.45	0.44	0.47
	e_6	0.71	0.59	0.66	0.69	0.57	0.68	0.71	0.62	0.62	0.58	0.56	0.66
	e_7	0.69	0.80	0.79	0.74	0.79	0.68	0.71	0.76	0.59	0.55	0.53	0.63
P_4	e_9	0.63	0.59	0.62	0.60	0.60	0.55	0.62	0.59	0.59	0.65	0.63	0.63
	e_{12}	0.63	0.59	0.57	0.55	0.60	0.55	0.65	0.62	0.59	0.55	0.53	0.63
	e_{13}	0.65	0.80	0.81	0.74	0.81	0.66	0.68	0.76	0.53	0.48	0.47	0.56
	e_{16}	0.75	0.75	0.74	0.71	0.76	0.74	0.71	0.73	0.65	0.58	0.56	0.63
	e_{18}	0.69	0.77	0.79	0.76	0.79	0.68	0.71	0.73	0.59	0.48	0.50	0.63
	e_{24}	0.71	0.43	0.53	0.52	0.43	0.58	0.53	0.43	0.68	0.61	0.56	0.59
	e_{30}	0.67	0.82	0.81	0.76	0.81	0.68	0.71	0.78	0.59	0.55	0.53	0.59

Tabela 4.23 Matriz de blocos para os valores de máximo do conjunto de dados City temperatures normalizado com 4 grupos de variáveis

		H_1	H_2				H_3			H_4			
		Mai.	Jan.	Fev.	Mar.	Dez.	Abr.	Out.	Nov.	Jun.	Jul.	Ago.	Set.
	e_{19}	0.69	0.93	0.94	0.88	0.93	0.79	0.74	0.86	0.56	0.48	0.47	0.59
	e_{20}	0.73	0.80	0.81	0.79	0.81	0.76	0.74	0.81	0.65	0.61	0.56	0.66
P1	e_{23}	0.63	0.86	0.87	0.79	0.81	0.68	0.71	0.76	0.47	0.42	0.41	0.59
	e_{32}	0.58	0.98	0.96	0.81	0.98	0.66	0.65	0.84	0.35	0.26	0.28	0.47
	e_1	0.52	0.39	0.38	0.48	0.36	0.45	0.44	0.41	0.44	0.39	0.47	0.47
	e_8	0.50	0.34	0.36	0.31	0.36	0.32	0.35	0.32	0.44	0.45	0.41	0.41
	e_{10}	0.73	0.50	0.66	0.60	0.50	0.68	0.65	0.51	0.74	0.77	0.72	0.69
	e_{11}	0.52	0.41	0.45	0.40	0.40	0.39	0.38	0.35	0.35	0.52	0.47	0.44
	e_{15}	0.52	0.43	0.47	0.43	0.43	0.39	0.41	0.41	0.44	0.45	0.41	0.44
P2	e_{21}	0.54	0.16	0.06	0.19	0.14	0.26	0.24	0.14	0.53	0.52	0.44	0.34
	e_{22}	0.54	0.32	0.38	0.40	0.31	0.42	0.38	0.32	0.47	0.48	0.47	0.47
	e_{26}	0.56	0.45	0.47	0.48	0.40	0.47	0.47	0.41	0.50	0.52	0.50	0.50
	e_{31}	0.48	0.18	0.19	0.14	0.31	0.26	0.26	0.24	0.41	0.45	0.38	0.31
	e_{35}	0.54	0.27	0.30	0.29	0.29	0.34	0.41	0.59	0.56	0.61	0.56	0.53
	e_{36}	0.56	0.32	0.38	0.38	0.33	0.42	0.38	0.32	0.50	0.52	0.47	0.44
	e_{37}	0.73	0.50	0.64	0.62	0.45	0.61	0.65	0.65	0.74	0.74	0.53	0.56
	e_2	0.69	0.57	0.57	0.57	0.60	0.55	0.68	0.62	0.71	0.77	0.75	0.72
	e_{14}	0.60	0.59	0.62	0.57	0.60	0.53	0.62	0.59	0.56	0.58	0.59	0.59
	e_{17}	0.67	0.50	0.57	0.57	0.48	0.55	0.59	0.51	0.71	0.84	0.78	0.72
	e_{25}	0.63	0.39	0.40	0.40	0.40	0.45	0.56	0.46	0.65	0.68	0.38	0.59
P3	e_{27}	0.65	0.55	0.60	0.57	0.55	0.55	0.62	0.57	0.68	0.74	0.72	0.69
	e_{28}	0.56	0.59	0.62	0.60	0.60	0.53	0.65	0.62	0.47	0.45	0.44	0.56
	e_{29}	0.63	0.45	0.45	0.38	0.45	0.47	0.71	0.65	0.59	0.74	0.69	0.72
	e_{33}	0.69	0.41	0.49	0.55	0.26	0.53	0.59	0.46	0.74	0.97	1.00	0.78
	e_{34}	0.65	0.50	0.53	0.50	0.55	0.53	0.62	0.57	0.59	0.68	0.72	0.69
	e_3	0.83	0.73	0.72	0.74	0.76	0.76	0.91	0.84	0.85	0.90	0.88	0.91
	e_4	0.85	0.93	0.91	0.90	0.98	0.89	0.94	1.00	0.79	0.71	0.69	0.78
	e_5	0.85	0.75	0.79	0.79	0.74	0.82	0.91	0.84	0.88	0.90	0.84	0.88
	e_6	0.92	0.91	0.94	1.00	0.88	1.00	0.94	0.92	0.82	0.77	0.75	0.84
	e_7	0.81	0.98	0.96	0.93	0.98	0.87	0.85	0.92	0.74	0.68	0.66	0.78
P4	e_9	0.88	0.82	0.83	0.86	0.88	0.87	1.00	0.95	0.91	1.00	0.97	1.00
	e_{12}	0.73	0.68	0.66	0.64	0.71	0.66	0.79	0.76	0.71	0.71	0.69	0.75
	e_{13}	0.83	1.00	1.00	0.98	1.00	0.92	0.91	0.97	0.79	0.74	0.75	0.84
	e_{16}	0.98	0.98	0.98	0.98	0.95	0.97	0.94	0.95	0.97	0.90	0.84	0.91
	e_{18}	0.81	0.91	0.89	0.88	0.90	0.87	0.85	0.89	0.76	0.68	0.63	0.72
	e_{24}	1.00	0.77	0.83	0.88	0.81	1.00	1.00	0.89	1.00	0.87	0.81	0.91
	e_{30}	0.79	0.98	0.96	0.93	0.98	0.87	0.88	0.95	0.74	0.71	0.69	0.78

CAPÍTULO 5

Conclusões

Nesse trabalho fizemos uma análise do algoritmo Block Clustering K-means para dados quantitativos e propomos o algoritmo de Block Clustering K-means para variáveis do tipo intervalo com dissimilaridade calculada pela distância Euclidiana.

Primeiramente nós fizemos uma análise do algoritmo de Block clustering aplicada a dados quantitativos, então fazendo diversas simulações em ambientes controlados vimos que o particionamento dos grupos de variáveis é feito de acordo com a homogeneidade dos valores dos indivíduos nas variáveis e há dependência em relação a quantidade de grupos de indivíduos disponíveis, pois quanto mais próximo a quantidade de grupos de indivíduos for da quantidade de classes à priori a homogeneidade pode ser distribuída de forma mais precisa nos blocos disponíveis. Assim, as variáveis dos indivíduos de determinada classe que possuem valores com pouca variação entre si, são distribuídas de forma mais precisa, enquanto quanto maior for essa variação a formação dos grupos será menos precisa. Em seguida, ainda aplicando o algoritmo BCKM a dados quantitativos, foram utilizadas bases de dados reais e comparamos a performance deste algoritmo em relação ao algoritmo BC, assim percebemos que com o aumento da quantidade de grupos de variáveis disponíveis, as variáveis vão se redistribuindo de acordo com sua proximidade em relação às outras variáveis, até que quando a quantidade de grupos de variáveis é igual a quantidade de variáveis, cada variável fica em grupos distintos.

Para o algoritmo de Block Clustering K-means para variáveis de tipo intervalo, proposto nesse trabalho, fizemos testes com bases de dados reais e comparando a performance, pelos índices OERC e CR, com o modelo KM percebemos que ela não é degredada dependendo da quantidade de grupos de variáveis escolhida para uma quantidade fixa de grupos de indivíduos e pode até trazer melhor desempenho, como no caso da base de dados Freshwater Fish para uma configuração com 2 clusters de variáveis. Com isso, dado que no algoritmo KM só obtemos a partição de indivíduos, com o algoritmo BCKM temos a inclusão da informação da partição de variáveis proposta pelo sem haver degradação ou, dependendo da quantidade de grupos de variáveis, com uma mínima degradação ou melhoria da performance. Por fim, do mesmo modo como acontece na aplicação do algoritmo BCKM para dados quantitativos, quando temos variáveis do tipo intervalo com uma quantidade fixa de grupos de indivíduos, as variáveis vão se adaptando à quantidade de grupos de variáveis disponíveis, assim com o aumento desta há uma tendência de espalhamento das variáveis de forma que para uma quantidade de grupos de variáveis igual a quantidade de variáveis, cada variável fica individualmente em um grupo e o algoritmo BCKM se comporta como o algoritmo K-means tradicional.

Como trabalhos futuros seria interessante a aplicação do método de agrupamento simultâneo para grupos de indivíduos e variáveis em agrupamentos Fuzzy, onde os dados não estão rigidamente ligados a um grupo, mas existe um grau de pertinência do indivíduo a cada grupo.

Referências Bibliográficas

[1] H. H. Bock. Clustering algorithms and kohonen maps for symbolic data. *ICNCB*, 15:1–13, 2002.

[2] H. H. Bock and E. Diday. *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Springer, Berlin Heidelberg, 2000.

[3] F. A. T. de Carvalho, P. Brito, and H. H. Bock. Dynamic clustering for interval data based on l2 distance. *Computational Statistics*, 21:231–250, 2006.

[4] M. R. P. Ferreira and F. de Carvalho. Kernel fuzzy c-means with automatic variable weighting. *Fuzzy Sets and Systems*, 2013.

[5] G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36:463–473, 2003.

[6] G. Govaert and M. Nadif. An em algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:643–647, 2005.

[7] G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52:233–3245, 2008.

[8] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition*, pages 651–666, 2010.

[9] M.Chavent, F. de Carvalho, Y. Lechevallier, and R. Verde. New clustering methods for inteval data. *Computational Statistics*, 21:211–229, 2006.

[10] J. van Rosmalen, P. J. F. Groenen, J. Trejos, and W. Castillo. Optimization strategies for two-mode partitioning. *Journal of Classification*, pages 155–181, 2009.