CLUSTERIZAÇÃO AUTOMÁTICA NA REDUÇÃO DA DIMENSIONALIDADE DOS DADOS

Éldman de Oliveira Nunes
Escola de Administração do Exército
Rua Território do Amapá, 455, 41.540-830, Salvador, BA, Brasil
eldman.nunes@gmail.com

Aura Conci
Universidade Federal Fluminense
Rua Passo da Pátria, 156, 24.210-240, Niterói, RJ, Brasil
aconci@ic.uff.br

Resumo

A clusterização é uma técnica extremamente importante para análise do comportamento dos dados e têm sido largamente utilizada para solução de diversos problemas de ordem prática. Entretanto, o desconhecimento do número ideal de grupos para partição da base, bem como o excesso de atributos que compõe os dados, são fatores que degradam a qualidade dos resultados. Este artigo apresenta uma nova estratégia para reduzir a dimensionalidade dos dados a partir da determinação automática do número ideal de grupos presentes em uma base. Os resultados obtidos com a nova estratégia proposta demonstram ser possível melhorar a abrangência e a acurácia na classificação dos dados.
Palavras-Chaves: Análise de Agrupamentos, segmentação, K-Means.

Abstract

The clustering is an extremely important technique for analysis of the behavior of the data that has been broadly used for solution of varied problems of practical order. However, the unknowledge of the ideal number of clusters for partition of the base, as well as the excess of attributes that composes the data are factors that degrade the quality of the results. This article presents one new strategy to reduce the dimensionality of the data from of the automatic determination of the ideal number of clusters present at the base. The obtained results with the new proposed strategy demonstrate to be possible to improve the abrangency and the accuracy in the classification of the data.
Keywords: Clusters Analysis, segmentation, K-Means

1. INTRODUÇÃO

A clusterização é uma ferramenta útil para o estudo e compreensão do comportamento de dados que tem sido empregada na solução de diversos problemas nas mais variadas áreas do conhecimento. Seu objetivo é organizar os objetos de uma base de dados em grupos de tal forma que os objetos dentro de um mesmo grupo sejam mais similares entre si do que com objetos de outros grupos.

Quando a solução do problema de agrupamento depende da informação *a priori* do número de grupos desejados para partição dos dados, este é referenciado na literatura como "problema de k-clusterização" [1]. Entretanto, na maioria das aplicações práticas, o número ideal de grupos é desconhecido e sua obtenção acrescenta complexidade na solução, sendo referenciado na literatura como "problema de clusterização automática" [2].

Anais do XI Simpósio de Pesquisa Operacional da Marinha (SPOLM 2008), Rio de Janeiro: Centro de Análises de Sistemas Navais, p. 1-14, 2008

Em ambos os casos, a eficiência da solução dependerá da regularidade e da uniformidade dos dados disponíveis. Dificilmente, porém, tal comportamento é encontrado na maioria das situações reais, pois estas envolvem uma grande massa de dados, compostos por diversos atributos. Considerando que a multidimensionalidade dos atributos usados para organizar os dados constitui-se em um fator complicador para solução de um problema, a seleção dos atributos mais relevantes do conjunto de atributos disponíveis na base de dados é um passo importante para diminuir o tempo de processamento, reduzir o espaço de busca e evitar erros de classificação.

A determinação automática do número ideal de grupos em um problema de clusterização e a redução da dimensionalidade dos atributos usados para organizar os dados são aspectos de grande utilidade na análise de agrupamentos e constitui-se no objetivo deste trabalho.

2. CLUSTERIZAÇÃO

Dado um conjunto X, um problema de Clusterização (análise de agrupamento) consiste em se agrupar os objetos (elementos) de X de modo que objetos mais similares fiquem no mesmo grupo (cluster) e os objetos menos similares sejam alocados para grupos distintos (figura 1).

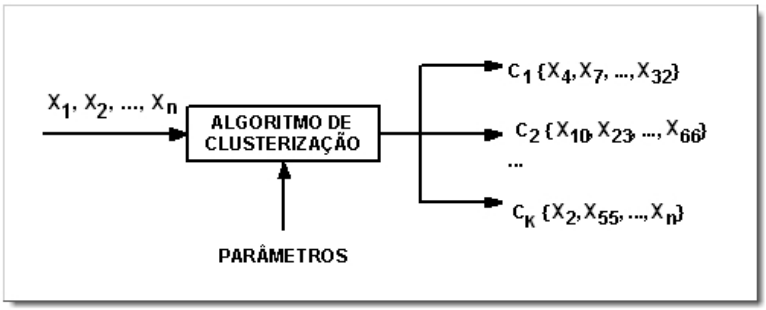

FIGURA 1 - Esquema funcional de um algoritmo de clusterização

Problemas de clusterização podem ser definidos da seguinte forma [3]:

Dado um conjunto com n elementos $X = \{X_1, X_2, ..., X_n\}$, o problema de clusterização consiste na obtenção de um conjunto de k clusters, $C = \{C_i, C_2, ..., C_k\}$, tal que haja uma maior similaridade entre os elementos contidos em um cluster C_i do que qualquer um destes com os elementos de um dos demais clusters do conjunto C. O conjunto C é considerado uma clusterização com k clusters caso as condições das equações (1-3) sejam satisfeitas:

$$\bigcup_{i=1}^{k} C_i = X \tag{1}$$

$$C_i \neq \emptyset, para: 1 \leq i \leq k \tag{2}$$

$$C_i \cap C_j = \emptyset, para: 1 \leq i, j \leq k, i \neq j \tag{3}$$

A busca pela melhor solução no espaço de soluções possíveis torna o processo de clusterização um problema NP-Difícil. A utilização de métodos exatos para obtenção da solução ótima fica impraticável, uma vez que a verificação exaustiva de todas as configurações de agrupamentos possíveis é computacionalmente inviável [3].

Para o problema de k-clusterização, quando se conhece o número de k *a priori*, existem N diferentes maneiras de agrupar n elementos em k clusters:

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \quad (4)$$

O que torna exponencial o crescimento de n. Por exemplo, existem os seguintes números de soluções possíveis para se combinar 10 elementos, 100 elementos e 1000 elementos em 2 clusters, respectivamente: 512 soluções, $6{,}33825\times10^{29}$ soluções e $5{,}3575\times10^{300}$ soluções.

Para o problema de clusterização automática, o número de combinações possíveis é dado pela equação:

$$N(n,k) = \sum_{k=1}^{n}\frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \quad (5)$$

Assim, aumenta significativamente o número de combinações possíveis. Para clusterização de um conjunto com apenas 10 elementos, em um número de clusters variável de 1 a 10, existem 115.975 maneiras diferentes.

2.1. MÉTODOS DE PARTICIONAMENTO

A fim de reduzir a complexidade na solução do problema, utilizam-se métodos heurísticos capazes de fornecer soluções sub-ótimas em tempo satisfatório. Na literatura são encontrados diversos desses métodos [4], a escolha dentre eles dependerá do propósito de cada aplicação particular e do tipo de dados disponíveis. Basicamente, esses métodos podem ser classificados em hierárquicos e de particionamento [1].

Os Métodos de Partição Baseados em Recolocação tem por objetivo particionar um conjunto de dados com n elementos em k grupos distintos de forma a minimizar um critério escolhido. Estes métodos tentam descobrir novos clusters realocando interativamente pontos entre os subconjuntos de forma a melhorar os clusters gradualmente, o que não ocorre nos métodos hierárquicos.

Freqüentemente, os métodos de particionamento encontram clusters com qualidade superior (maior similaridade interna) aos dos encontrados pelos métodos hierárquicos [5]. Devido a este melhor desempenho, os algoritmos de particionamento normalmente são mais empregados. Dentre eles, os que são baseados em um ponto central (k-means) [6] ou os que são baseados em um objeto representativo para o cluster (k-medoids) [7].

A qualidade do resultado obtido com os métodos de particionamento depende da seleção coerente das seguintes variáveis: seleção do algoritmo de agrupamento e dos atributos relevantes, escolha do número de grupos e medidas de similaridade, definição dos critérios de agrupamento e homogeneização das variáveis.

2.2. MEDIDAS DE SIMILARIDADE

O objetivo da clusterização é agrupar os elementos de um conjunto tomando por base o seu grau de similaridade. Para medir o quanto um elemento é similar a outro, a fim de determinar se deve pertencer ou não a um mesmo cluster, são utilizadas medidas de similaridade. O critério de similaridade mais comum quando se utilizam atributos numéricos baseia-se nas funções de distância.

Para empregar estas funções é preciso representar cada elemento como um vetor no espaço n dimensional das características. Neste caso, quanto menor for à distância entre um par de elementos, maior é a similaridade entre eles. Espera-se que à distância entre objetos de um mesmo agrupamento seja significativamente menor do que à distância entre objetos de

agrupamentos diferentes. Dentre as medidas de distância mais utilizadas encontram-se a distância "city-block " (para $r=1$) e a distância euclidiana (para $r=2$), conforme a equação:

$$d(X_i, X_j) = \left[\sum_{l=1}^{n} \left| x_{il} - x_{jl} \right|^r \right]^{\frac{1}{r}} \quad (6)$$

2.3. MÉTODO K-MEANS

K-means é um método amplamente difundido, existindo muitas variações propostas na literatura e diversos nomes (K-médias, isodata, ou migração de médias). K-means é um método de partição baseado em recolocação que necessita da definição *a priori* do número de agrupamentos k.

O critério de custo a ser minimizado é definido em função da distância dos elementos em relação aos centros dos agrupamentos. Usualmente, este critério é a soma residual dos quadrados das distâncias (geralmente é usada a distância euclidiana). Entende-se por soma residual dos quadrados, a soma dos quadrados das distâncias dos elementos ao centróide do seu cluster, conforme a equação:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - X_i \right) \left(x_{ij} - X_i \right)^t \quad (7)$$

Onde x_{ij} é o j-ésimo objeto do cluster i, X_i é o representante do cluster i (a média ou mediana dos objetos do cluster), e n_i é a quantidade de objetos do cluster i.

O elemento representativo de um cluster é o seu centróide, que possui um valor médio para os atributos considerados, relativos a todos os elementos do cluster. A utilização do centróide como elemento representativo de um cluster é conveniente apenas para atributos numéricos e possui um significado geométrico e estatístico claro, podendo, entretanto, receber mais influência de um único elemento que se encontre próximo à fronteira do cluster.

A partir de uma estimativa inicial das coordenadas dos centros dos agrupamentos (centróides), o algoritmo calcula a distância de cada ponto do conjunto a estes centróides. A seguir, o algoritmo aloca cada elemento do conjunto em um grupo, de acordo com a menor distância ao centróide correspondente. A nova estimativa das coordenadas dos centróides é calculada pela média aritmética das coordenadas dos pontos associados a cada grupo (figura 2).

O método K-means é sensível ao particionamento inicial realizado, em virtude da escolha das coordenadas dos k centróides dos clusters ser feita inicialmente de forma aleatória. A partir deste primeiro particionamento, o algoritmo realiza uma busca de um ponto de máximo para o seu critério de parada. Não há garantias de que o algoritmo encontre o máximo global, sendo possível encontrar soluções distintas em diferentes execuções do algoritmo.

O K-means é um tipo de algoritmo normalmente utilizado para a classificação não-supersionada. Este algoritmo é constituído dos seguintes passos básicos:

(1) Determinar as posições iniciais dos k centróides dos clusters;

(2) Alocar cada elemento ao cluster do centróide mais próximo;

(3) Recalcular os centros dos clusters a partir dos elementos alocados;

(4) Repetir os passos de 2 a 4 segundo algum critério de convergência

Como critério de convergência pode se executar o algoritmo até que os centróides não se movam mais ou até que um determinado número máximo de interações seja alcançado.

.

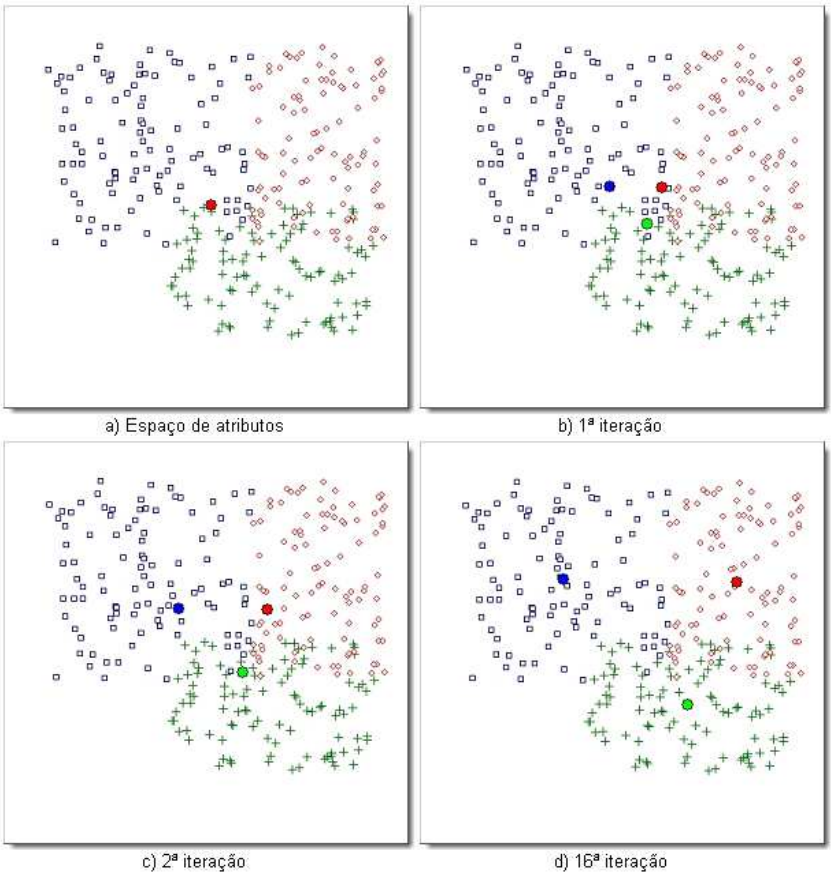

FIGURA 2 - Aplicação do método K-Means

3. METODOLOGIA

Existem algumas técnicas para clusterização automática baseada em junções ou divisões sucessivas (métodos hierárquicos), noções de densidade, variação do nível de quantização do espaço, modelos matemáticos e estatísticos. Apesar destas técnicas obterem bons resultados, em geral apresentam elevado custo computacional [8].

Um critério para determinar a qualidade da clusterização é apresentado em [9] e estabelece a seguinte função de k:

$$G(k) = \frac{(n-k)B}{(k-1)W} \tag{8}$$

Onde k é o número de clusters usados para a segmentação, n o número de objetos da base de dados, W conforme a equação (5), e B é a variação não explicada do somatório dos quadrados das diferenças de toda a base de dados para a média a partir do somatório dos quadrados das diferenças de cada objeto de cada cluster para o centróide do seu cluster, dado pela equação:

$$B = T - W = \sum_{i=1}^{k} n_i (X_i - X)(X_i - X)^t \tag{9}$$

T é a soma dos quadrados das diferenças de cada objeto da base de dados para a média de todos os objetos da base (X), conforme a equação:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - X)(x_{ij} - X)^t \tag{10}$$

O valor de $G(k)$ representa a estatística da análise da variância do agrupamento formado. Quanto maior o seu valor, mais homogêneos serão os objetos dentro de cada grupo e melhor será a partição.

Para determinação automática do número ideal de agrupamentos foi desenvolvido um sistema, baseado no algoritmo K-Means, que executa o particionamento de uma base de dados para cada um dos possíveis valores de k ($1 \pi k \pi n$) maximizando a equação (8).

Como o algoritmo K-Means possui o inconveniente de ter sua solução influenciada pela escolha da configuração inicial (eleição arbitrária dos k objetos como centros iniciais dos clusters), optou-se por inicializar os centróides de forma a dispersá-los uniformemente sobre o espaço de representação. Esta estratégia tende a fornecer melhores resultados, além de contribuir para a aceleração da convergência.

Para redução da dimensionalidade dos dados é necessário realizar a seleção dos atributos mais relevantes, ou seja, que mais contribuam para separabilidade das classes, dentro do conjunto de atributos disponíveis. Basicamente, este deve ser um processo de múltiplas iterações, consistindo dos seguintes passos:

(1) Selecionar um subconjunto de atributos (atributos candidatos) de acordo com algum critério;
(2) Aplicar o algoritmo sobre a base de dados com este subconjunto de atributos;
(3) Medir a qualidade da seleção, conforme equação (8);
(4) Repetir os passos 1 a 3 até que se encontre um resultado satisfatório.

4. EXPERIMENTOS

Para realização dos testes, foi selecionada a popular base de dados *Iris Plants Database*, a fim de testar todas as possíveis combinações de seus atributos. Esta base pode ser obtida no repositório de bases de dados para descoberta de conhecimento da Universidade da Califórnia, Irvine (*UCI MachineLearning Repository*) [10].

A base Íris consiste de 3 classes: Íris-Setosa, Íris-Virginica e Íris-Versicolor, cada uma possuindo 50 instâncias, com distribuição de 33.3% para cada uma das 3 classes. A classe Íris-Setosa é linearmente separável das outras duas; Já as classes Íris-Virginica e Íris-Versicolor não são linearmente separáveis. Cada instância possui cinco atributos, sendo quatro do tipo numérico (comprimento da sépala em cm, largura da sépala em cm, comprimento da pétala em cm, largura da pétala em cm) e um do tipo categórico (classe: Íris-Setosa, Íris-Versicolor, Íris-Virginica). O atributo do tipo categórico não participa do processo de clusterização, entretanto define a qual classe a instância pertence.

Para execução dos testes foi utilizado um PC com a seguinte configuração: Processador Intel Pentium 4, CPU 2.40 GHz, Memória DRAM 511 MB, Disco rígido de 80 GB, Sistema Operacional Microsoft Windows XP Professional.

Os experimentos foram divididos em três etapas. A primeira consistiu da clusterização da base de dados considerando todos os quatro atributos. A segunda, considerando três atributos e a terceira, considerando apenas dois atributos.

4.1. PRIMEIRA ETAPA: UTILIZAÇÃO DE TODOS OS ATRIBUTOS

A figura 3 apresenta a clusterização da base Íris considerando todas as suas 150 instâncias, cada uma com seus quatro atributos. Para este caso de teste, o valor de k variou entre 2 e 150. O tempo médio para realização desta clusterização foi de 14, 2 segundos.

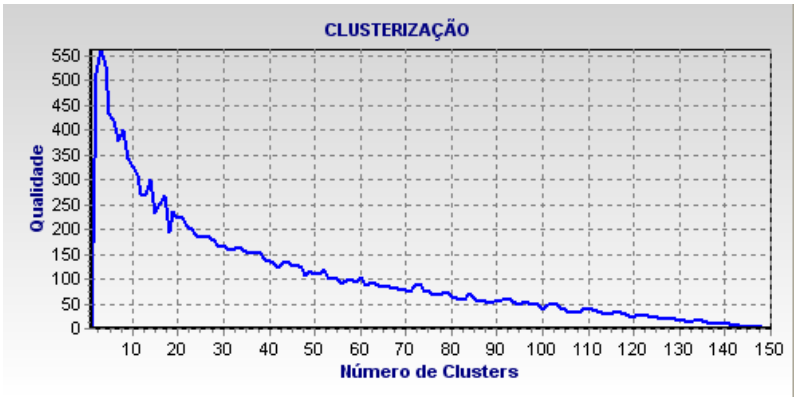

FIGURA 3 – Clusterização com k variando entre 2 e 150

A figura 4 apresenta a mesma clusterização com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,94 segundos. É possível constatar que o número de clusters encontrado corresponde ao número de classes existentes na base, ou seja, igual a três.

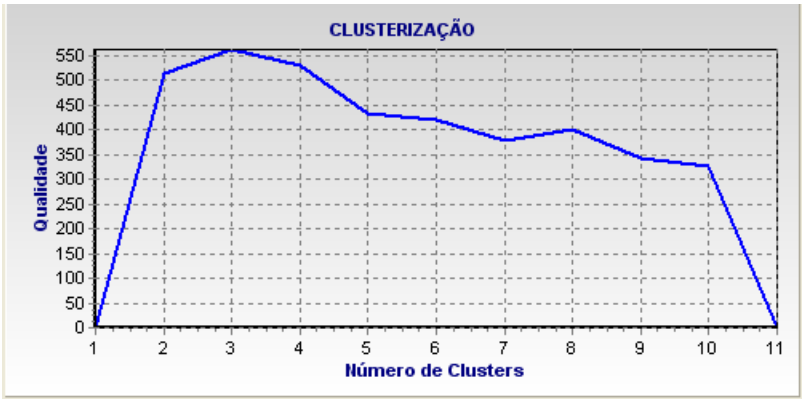

FIGURA 3 – Clusterização com k variando entre 2 e 10

A tabela 1 apresenta os resultados obtidos através da abrangência - equação (9), acurácia - equação (10), bem como da matriz de classificação, considerando a manutenção dos quatro atributos das instâncias. A matriz de classificação indica o número de amostras classificadas em cada classe. A situação ideal é estar toda a amostragem de uma classe na diagonal principal. Valores fora da diagonal principal indicam erro de classificação.

$$Abrangência = \frac{Número\ de\ Amostras\ da\ Classe\ no\ Cluster}{Número\ Total\ de\ Amostras\ da\ Classe} \quad (9)$$

$$Acurácia = \frac{Número\ de\ Amostras\ da\ Classe\ no\ Cluster}{Número\ Total\ de\ Amostras\ do\ Cluster} \quad (10)$$

TABELA 1 - Caso de teste: todos os quatro atributos

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
1	Iris-versicolor	94,00%	77,05%
2	Iris-virginica	72,00%	92,31%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	0	0	50
1	Iris-versicolor	0	47	14	61
2	Iris-virginica	0	3	36	39
TOTAL		50	50	50	150

Da análise da tabela 1 verifica-se que a classe Íris-Setosa é linermente separável, enquanto que as classes Íris-Vesicolor e Íris-Verginica não.

4.2. SEGUNDA ETAPA: REMOÇÃO DE UM ATRIBUTO

A segunda etapa de testes consistiu na remoção de um dos atributos. Foram realizados quatro testes, um para cada atributo removido, dos quatro possíveis.

4.2.1. Remoção do atributo comprimento da sépala

A tabela 2 apresenta os resultados obtidos com a remoção do atributo comprimento da sépala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,78 segundos.

TABELA 2 - Remoção do atributo comprimento da sépala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
1	Iris-versicolor	94,00%	92,16%
2	Iris-virginica	92,00%	93,88%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	0	0	50
1	Iris-versicolor	0	47	4	51
2	Iris-virginica	0	3	46	49
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 2 que a abrangência da classe Íris-Vesicolor se manteve constante em 94% e acurácia melhorou de 77,05% para 92,16%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 92% e a acurácia de 92,31% para 93,88%. O que indica que o conjunto de atributos (largura da sépala, comprimento da pétala e largura da pétala) é relevante para separabilidade das classes permitindo obter um resultado melhor do que com quatro atributos.

4.2.2. Remoção do atributo largura da sépala

A tabela 3 apresenta os resultados obtidos com a remoção do atributo largura da

sépala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,94 segundos.

TABELA 3 - Remoção do atributo largura da sépala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
1	Iris-versicolor	96,00%	80,00%
2	Iris-virginica	76,00%	95,00%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	0	0	50
1	Iris-versicolor	0	48	12	60
2	Iris-virginica	0	2	38	40
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 3 que a abrangência classe Íris-Vesicolor melhorou de 94% para 96% e acurácia de 77,05% para 80%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 76% e a acurácia de 92,31% para 95%. O que indica que o conjunto de atributos (comprimento da sépala, comprimento da pétala e largura da pétala) é relevante para separabilidade das classes permitindo encontrar um resultado melhor do que com quatro atributos.

4.2.3. Remoção do atributo comprimento da pétala

A tabela 4 apresenta os resultados obtidos com a remoção do atributo comprimento da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,94 segundos.

TABELA 4 - Remoção do atributo comprimento da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
1	Iris-versicolor	78,00%	72,22%
2	Iris-virginica	70,00%	76,09%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	0	0	50
1	Iris-versicolor	0	39	15	54
2	Iris-virginica	0	11	35	46
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 4 que a abrangência classe Íris-Vesicolor piorou de 94% para 78% e acurácia de 77,05% para 72,22%, enquanto na classe Íris-Virginica a abrangência piorou de 72% para 70% e a acurácia de 92,31% para 76,09%. O que indica que o conjunto de atributos (comprimento da sépala, largura da sépala e largura da pétala) não é relevante para separabilidade das classes.

4.2.4. Remoção do atributo largura da pétala

A tabela 5 apresenta os resultados obtidos com a remoção do atributo largura da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,79 segundos.

TABELA 5 - Remoção do atributo largura da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
1	Iris-versicolor	90,00%	77,59%
2	Iris-virginica	74,00%	88,10%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	0	0	50
1	Iris-versicolor	0	45	13	58
2	Iris-virginica	0	5	37	42
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 5 que a abrangência classe Íris-Vesicolor piorou de 94% para 90% e acurácia melhorou de 77,05% para 77,59%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 74% e a acurácia piorou de 92,31% para 88,10%. O que indica que o conjunto de atributos (comprimento da sépala, largura da sépala e comprimento da pétala) não é relevante para separabilidade das classes.

4.3. TERCEIRA ETAPA: REMOÇÃO DE DOIS ATRIBUTOS

A terceira etapa de testes consistiu na remoção de dois dos atributos. Foram realizados seis testes, correspondendo aos arranjos sem repetição possíveis.

4.3.1. Remoção dos atributos comprimento e largura da pétala

A tabela 6 apresenta os resultados obtidos com a remoção dos atributos comprimento da pétala e largura da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,78 segundos.

TABELA 6 - Remoção dos atributos comprimento e largura da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	67,57%
	Iris-versicolor	42,00%	28,38%
	Iris-virginica	6,00%	4,05%
1	Iris-setosa	0,00%	0,00%
	Iris-versicolor	58,00%	38,16%
	Iris-virginica	94,00%	61,84%

Cluster	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	50	21	3	74
1	0	29	47	76
TOTAL	50	50	50	150

Pode-se verificar na tabela 6 que apenas dois clusters foram formados. Um cluster teve abrangência de 100% da classe Íris-Setosa e o outro cluster teve abrangência de 94% da classe Íris-Virginica. As amostras da classe Íris-Vesicolor ficaram distribuídas entre estes dois clusters nas proporções de 42% e 58%, respectivamente. O que indica que o conjunto de atributos (comprimento da sépala e largura da sépala) não é relevante para separabilidade das classes.

4.3.2. Remoção dos atributos largura da sépala e comprimento da pétala

A tabela 7 apresenta os resultados obtidos com a remoção dos atributos largura da sépala e comprimento da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,78 segundos.

TABELA 7 - Remoção dos atributos largura da sépala e comprimento da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	94,34%
1	Iris-versicolor	76,00%	74,51%
2	Iris-virginica	74,00%	80,43%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	50	3	0	53
1	Iris-versicolor	0	38	13	51
2	Iris-virginica	0	9	37	46
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 7 que a abrangência classe Íris-Vesicolor piorou de 94% para 76% e acurácia piorou de 77,05% para 74,51%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 74% e a acurácia piorou de 92,31% para 80,43%. O que indica que o conjunto de atributos (comprimento da sépala e largura da pétala) não possui relevância para separabilidade das classes.

4.3.3. Remoção dos atributos comprimento da sépala e da pétala

A tabela 8 apresenta os resultados obtidos com a remoção dos atributos comprimento da sépala e comprimento da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,62 segundos.

TABELA 8 - Remoção dos atributos comprimento da sépala e da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	98,00%	100,00%
1	Iris-versicolor	92,00%	86,79%
2	Iris-virginica	88,00%	91,67%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	Iris-setosa	49	0	0	49
1	Iris-versicolor	1	46	6	53
2	Iris-virginica	0	4	44	48
TOTAL		50	50	50	150

Pode-se verificar na tabela 1 e na tabela 8 que a abrangência da classe Íris-Setosa piorou de 100% para 98%, a abrangência classe Íris-Vesicolor piorou de 94% para 92% e acurácia melhorou de 77,05% para 86,79%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 88% e a acurácia piorou de 92,31% para 91,67%. O que indica que o conjunto de atributos (largura da sépala e largura da pétala) possui relevância para separabilidade das classes permitindo obter um resultado satisfatório com apenas dois atributos.

4.3.4. Remoção dos atributos largura da sépala e da pétala

A tabela 9 apresenta os resultados obtidos com a remoção dos atributos largura da sépala e largura da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,62 segundos.

TABELA 9 - Remoção dos atributos largura da sépala e da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	0,00%	0,00%
1	Iris-setosa	0,00%	0,00%
	Iris-versicolor	54,00%	96,43%
	Iris-virginica	2,00%	3,57%
2	Iris-setosa	0,00%	0,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	44,00%	100,00%
3	Iris-setosa	0,00%	0,00%
	Iris-versicolor	46,00%	46,00%
	Iris-virginica	54,00%	54,00%

Cluster	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	50	0	0	50
1	0	27	1	28
2	0	0	22	22
3	0	23	27	50
TOTAL	50	50	50	150

Pode-se verificar na tabela 9 que quatro clusters foram formados. O primeiro cluster teve abrangência de 100% da classe Íris-Setosa com acurácia de 100%, o segundo cluster teve abrangência de 54% da classe Íris-vesicolor com acurácia de 96,43%, o terceiro cluster teve abrangência de 44% classe Íris-Virginica com acurácia de 100%. As demais amostras das classes Íris-Vesicolor e Íris-Virginica ficaram distribuídas no quarto cluster nas proporções de 46% e 54%, respectivamente. O que indica que o conjunto de atributos (comprimento da sépala e comprimento da sépala) possui relativa relevância para separabilidade das classes.

4.3.5. Remoção dos atributos comprimento e largura da sépala

A tabela 10 apresenta os resultados obtidos com a remoção dos atributos comprimento da sépala e largura da sépala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,47 segundos.

TABELA 10 - Remoção dos atributos comprimento e largura da sépala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	0,00%	0,00%
1	Iris-setosa	0,00%	0,00%
	Iris-versicolor	56,00%	96,55%
	Iris-virginica	2,00%	3,45%
2	Iris-setosa	0,00%	0,00%
	Iris-versicolor	34,00%	100,00%
	Iris-virginica	0,00%	0,00%
3	Iris-setosa	0,00%	0,00%
	Iris-versicolor	10,00%	16,13%
	Iris-virginica	52,00%	83,87%
4	Iris-setosa	0,00%	0,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	46,00%	100,00%

Cluster	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	50	0	0	50
1	0	28	1	29
2	0	17	0	17
3	0	5	26	31
4	0	0	23	23
TOTAL	50	50	50	150

Pode-se verificar na tabela 10 que cinco clusters foram formados. O primeiro cluster teve abrangência de 100% da classe Íris-Setosa com acurácia de 100%, o segundo cluster

teve abrangência de 56% da classe Íris-Vesicolor com acurácia de 96,55%, o terceiro cluster teve abrangência de 34% da classe Íris-Vesicolor com acurácia de 100%. O segundo e o terceiro clusters juntos cobrem 90% das amostras da classe Íris-Vesicolor com 90% de acurácia. O quarto cluster teve abrangência de 52% classe Íris-Virginica com acurácia de 83,87%, o quinto cluster teve abrangência de 46% classe Íris-Virginica com acurácia de 100%. O quarto e o quinto clusters juntos cobrem 98% da classe Íris-Virginica com 98% de acurácia. O que indica que o conjunto de atributos (comprimento da pétala e largura da sépala) possui relevância relativa para separabilidade das classes.

4.3.6. Remoção dos atributos comprimento da sépala e largura da pétala

A tabela 11 apresenta os resultados obtidos com a remoção dos atributos comprimento da sépala e largura da pétala. A clusterização foi realizada com o valor de k variando entre 2 e 10. O tempo médio para realização desta clusterização foi de 0,62 segundos.

TABELA 11 - Remoção dos atributos comprimento da sépala e largura da pétala

Cluster	Classe	Abrangência	Acurácia
0	Iris-setosa	100,00%	100,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	0,00%	0,00%
1	Iris-setosa	0,00%	0,00%
	Iris-versicolor	48,00%	100,00%
	Iris-virginica	0,00%	0,00%
2	Iris-setosa	0,00%	0,00%
	Iris-versicolor	0,00%	0,00%
	Iris-virginica	62,00%	100,00%
3	Iris-setosa	0,00%	0,00%
	Iris-versicolor	52,00%	57,78%
	Iris-virginica	38,00%	42,22%

Cluster	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
0	50	0	0	50
1	0	24	0	24
2	0	0	31	31
3	0	26	19	45
TOTAL	50	50	50	150

Pode-se verificar na tabela 11 que quatro clusters foram formados. O primeiro cluster teve abrangência de 100% da classe Íris-Setosa com acurácia de 100%, o segundo cluster teve abrangência de 48% da classe Íris-vesicolor com acurácia de 100%, o terceiro cluster teve abrangência de 62% classe Íris-Virginica com acurácia de 100%. As demais amostras das classes Íris-Vesicolor e Íris-Virginica ficaram distribuídas no quarto cluster nas proporções de 52% e 38%, respectivamente. O que indica que o conjunto de atributos (comprimento da sépala e comprimento da sépala) possui relevância relativa para separabilidade das classes.

5. CONCLUSÃO

A determinação automática do número ideal de clusters de uma base, bem como a redução da dimensionalidade dos dados através da seleção dos atributos mais relevantes para separabilidade das classes são duas funções de grande utilidade para análise do comportamento dos dados.

Este trabalho combinou as seguintes estratégias para solução do problema de clusterização automática: implementação do método K-Means, inicialização dos centróides com dispersão uniforme sobre o espaço de representação e maximização do valor da função G(k).

A fim de avaliar os resultados, realizou-se a clusterização da base de dados *Íris Plants Database* utilizando-se um processo manual de múltiplas iterações para seleção de um

conjunto de atributos candidatos relevantes.

Para o conjunto de três atributos, os melhores resultados foram, em ordem de prioridade: {largura da sépala, comprimento da pétala e largura da pétala} (seção 4.2.1), seguido do conjunto {comprimento da sépala, comprimento da pétala e largura da pétala} (seção 4.2.2). Os resultados dos testes comprovaram ser possível melhorar o desempenho da clusterização com a redução de um atributo.

Para o conjunto de dois atributos, os melhores resultados foram, em ordem de prioridade: {largura da sépala e largura da pétala} (seção 4.3.3), seguido do conjunto {comprimento da pétala e largura da sépala} (seção 4.3.5). Os resultados revelaram que é possível encontrar resultados satisfatórios mesmo com apenas dois atributos.

Os diferentes arranjos possíveis de 4 atributos resultaram em 10 casos de testes. Para cinco atributos são necessários 25 casos de testes. Devido a sua natureza combinatorial, a solução proposta para bases com grande número de atributos torna-se um problema NP-Dificil ou intratável computacionalmente. Em trabalhos futuros utilizar-se-á uma estratégia metaheurística híbrida.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] FASULO, D. An analysis of recent work on clustering algorithms. Technical report, 01-03-02, Seattle: University Of Washington - Department Of Computer Science And Engineering, 1999. 23 p.

[2] DOVAL, D.; MANCORIDIS, S.; MITCHELL, B.S. Automatic Clustering of Software Systems Using a Genetic Algorithm. step, p. 73, Software Technology and Engineering Practice, 1999.

[3] OCHI, L.S. Problemas de Clusterização em Mineração de Dados. In: Encontro Regional de Informática RJ/ES, 2004, Vitória. Anais do ERI 2004 RJ/ES, Vitória: ERI, 2004. p.1-6. CD-ROM.

[4] BERKHIN, P. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, California, 2002.

[5] NG, R.T.; HAN, J. Efficient and Effective Clustering Methods for Spatial Data Mining, In: International Conference on Very Large Data Base, 20, 1994, Santiago. Proceedings of DBLP. Santiago: Morgan Kaufmann,1994. p. 144-155.

[6] ZHANG, B.; HSU, M.; DAYAL,U. K-harmonic Means - A Spatial Clustering Algorithm with Bossting. In: International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, 1, 2000. Berlin. Lecture Notes in Artificial Intelligence. London: Springer-Verlag, Eds. J. F. Roddick & K. Hornsby, 2000. v. 2007, p.31-45.

[7] KAUFMAN, L.; ROUSSEEUW, P.J. Finding Groups in Data: an Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990.

[8] HAN, J.; KAMBER, M. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann, 2001. 550 p.

[9] CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, v. 3, (1), p. 1-27, 1974.

[10] ASUNCION, A.; NEWMAN, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.