Workshop de Trabalhos de Iniciação Científica e Graduação Bahia, Alagoas e Sergipe (WTICG)/VIII Escola Regional de Computação dos Estados da Bahia, Alagoas e Sergipe (ERBASE), Salvador, 2008. http://www.erbase2008.ufba.br/

Clusterização K-Means: Uma Proposta de Melhoria

Gustavo do S. Cardoso[1], Victor L. R. Nascimento[1], Eldman de Oliveira Nunes[1], Ernesto de Souza Massa Neto[1]

[1]Faculdade Hélio Rocha (FHR) – Salvador – BA – Brasil

{gcardoso, vrnascimento}@gmail.com, eldman@bol.com.br, massa.ernesto@gmail.com

Abstract. *The preparation of a method to obtain superior results to that one's found by the k-means's clustering technique using its own methodology as base consists on the objective of this study. With the aim of proving the obtained results by this study, the bases of data Iris Flower, Vote and Heart Disease were selected, which ones are quite known in the literature. These bases of data are essential to provide the necessary support and required validation for the tests. This study can represent an advancement in the scientific knowledge as it concerns the process of clustering of the datas without supervision. In this work, the results of the tests were accomplished and so have reached its main goal.*

Key-words: *Cluster, k-means, Similarity's Measure, Weight.*

Resumo. *Elaborar um método para obter resultados superiores aos resultados encontrados pela técnica de clusterização k-means utilizando sua própria metodologia como base é o objetivo deste estudo. A fim de comprovar os resultados obtidos por este estudo foram selecionadas as bases de dados Iris Flower, Vote e Heart Disease, as quais são bastante conhecidas na literatura, para fornecerem suporte e validação para os testes realizados. Este estudo pode significar um avanço no conhecimento científico no que se diz respeito ao processo de clusterização de dados não supervisionado, haja vista, os resultados dos testes realizados nesta pesquisa empírica foram satisfatórios, alcançando desta forma, o objetivo do estudo.*

Palavras-chave: *Cluster, k-means, Medida de Similaridade, Peso.*

1. Introdução

Os problemas de clusterização já são bastante conhecidos e estudados na literatura, principalmente nas áreas: da estatística e da matemática (BERKHIN, 2002). Entretanto, na área da computação estes problemas surgiram com a popularização do conceito das técnicas de mineração de dados (Data Mining) (OCHI, DIAS e SOARES, 2004).

Um problema de clusterização consiste em analisar e classificar um determinado conjunto de dados (BERKHIN, 2002), através do descobrimento de padrões grupais, que por sua vez, são descobertos utilizando técnicas que medem o nível de similaridade dos dados analisados.

Este estudo tem como foco propor uma melhoria no processo de clusterização da técnica k-means, a qual pertence ao grupo de clusterização particionada. Os métodos deste grupo buscam encontrar, iterativamente, a melhor partição dos *n* objetos em *k* grupos, onde o número de *k* partições é fornecido por um agente externo, este podendo ser o usuário de um sistema, bem como um algoritmo especialista, que tenha como finalidade encontrar a quantidade ideal de grupos a serem formados.

A pesquisa realizada neste estudo tem caráter empírico, necessitando que a comprovação dos resultados alcançados por ela seja realizada através de testes em um conjunto conhecido de bases de dados. Para tanto, as bases de dados selecionadas para a realização dos testes podem ser encontradas no repositório de bases de dados para descoberta de conhecimento da Universidade da Califórnia, Irvine (UCI Machine Learning Repository) (ASUNCION e NEWMAN, 2007).

O trabalho está organizado da seguinte forma. A sessão dois abrange o referencial teórico, mostrando o conceito de clusterização, como também as etapas necessárias para que um processo de clusterização ocorra. Na sessão três é feita uma abordagem sobre as bases de dados utilizadas para testes. A sessão quatro trata, de forma concisa, as hipóteses sobre a melhoria do algoritmo k-means. A quinta sessão mostra com maiores detalhes, os resultados obtidos com os testes realizados nas bases. E a sexta sessão apresenta as conclusões deste estudo e propostas futuras. A seguir, as referências bibliográficas que deram suporte ao estudo desenvolvido.

2. Clusterização

Clusterização é o processo de agrupamento de dados de forma não supervisionada, ou seja, sem a presença de uma entidade externa (supervisor) que defina previamente como os padrões de cada cluster devem ser gerados (JAIN e DUBES, 1988 apud PIMENTEL, VILMA e OMAR, 2003). A descoberta dos padrões existentes na base de dados é realizada através do próprio algoritmo de clusterização, que verifica a similaridade entre cada um deles e os agrupam, de forma a obter alta homogeneidade dentro dos grupos e alta heterogeneidade entre os grupos.

O processo clusterização representa uma das principais etapas do processo de análise de dados, denominada análise de clusters. A análise de cluster segundo Jain, Murty e Flynn (1999), é a tarefa de organizar uma coleção de dados que são normalmente representados como um vetor de medidas, ou por um ponto em um espaço multidimensional dentro de grupos baseados na similaridade entre eles. Grupos formados a partir de um processo de clusterização são comumente conhecidos por *clusters*.

Agrupamento de dados, tomada de decisões, segmentação de imagem (JAIN, MURTY e FLYNN, 1999 e CARLANTONIO, 2001), sistemas de recomendação, computação visual e gráfica, computação médica, diagnósticos médicos, biologia computacional, redes de comunicações, engenharia de transportes, redes de computadores e sistemas de manufatura (OCHI, DIAS e SOARES, 2002), são exemplos de áreas que podem ser exploras a partir do uso de técnicas de clusterização de dados.

2.1. Etapas de uma tarefa de clusterização

Jain, Murty e Flynn, (1999), definem de forma clara as etapas básicas envolvidas no processo de clusterização de dados. Estas etapas representam uma metodologia que perpassa todo o processo (Figura 1) de clusterização de dados, sendo necessárias para a realização deste processo as fases: representação dos dados (chamados também de objetos) (Figura 1-a); definição de uma medida de similaridade (Figura 1-b); realização do processo de clusterização (agrupamento) (Figura 1-c); e apresentação do resultado final (Figura 1-d).

A etapa de representação dos objetos (Figura 1-a) envolve a definição do número, do tipo e do modo de apresentação dos atributos que descrevem cada objeto. Neste momento serão definidos quais dados vão participar da clusterização em relação à totalidade da base de dados analisada. Esta etapa pode ser refinada fazendo uso de duas técnicas: a Seleção de Características e a Extração de Características. O resultado deste processo é uma nova representação dos objetos, exclusiva para o processo de clusterização. A primeira é um processo de identificação do subconjunto mais efetivo dos atributos disponíveis para descrever cada objeto, enquanto a segunda é o uso de uma ou mais transformações junto aos atributos de entrada de modo a salientar uma ou mais característica dentre aquelas que estão presentes nos dados. O resultado deste processo é uma nova representação dos objetos.

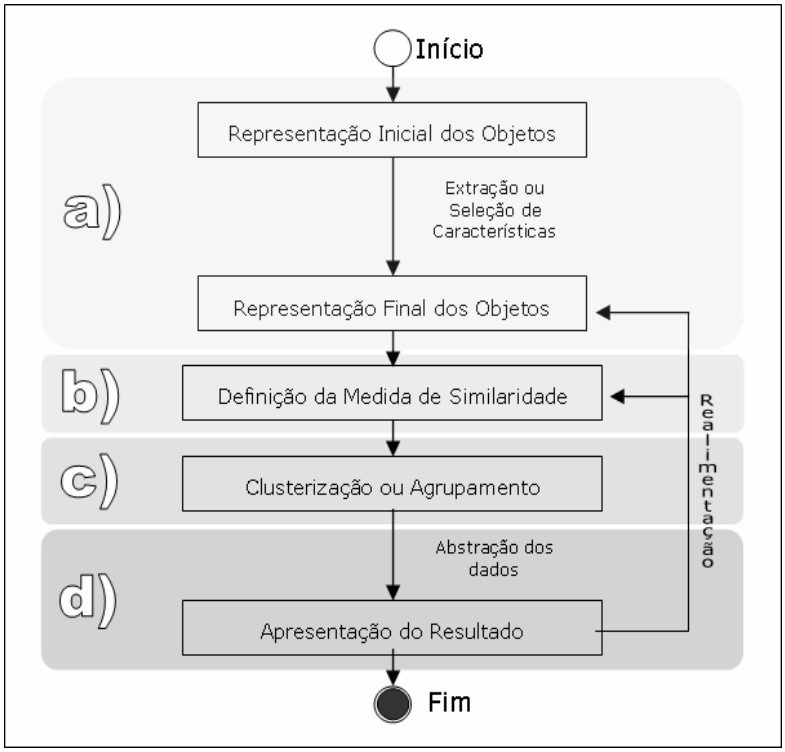

Figura 1: Etapas do Processo de Clusterização

Na etapa de definição da medida de similaridade (Figura 1-b) será necessário definir também uma técnica de agrupamento, como por exemplo: K-MEANS, K-MEDOIDS, DBSCAN, STING, ROCK, entre outras (BERKHIN, 2002). Já a medida de similaridade é fornecida por uma função de distância definida entre pares de dados ou

objetos, podendo ser: a medida de Minkowski, a medida de Chebyshev, a medida de Mahalanobis. É possível incluir na medida de distância aspectos conceituais (qualitativos) ou então numéricos (quantitativos).

A formação de clusters ou agrupamento (Figura 1-c) é a etapa que define como os grupos se os grupos serão organizados como conjuntos crisp (um objeto pertence ou não-pertence a um dado grupo) ou fuzzy (um padrão pode apresentar graus de pertinência aos grupos). O processo de agrupamento pode ser hierárquico, com um processo recursivo de junções ou separações de grupos, ou não-hierárquico, com o emprego direto de técnicas de discriminação de clusters.

Na etapa final validação (Figura 1-d), há a avaliação e a interpretação dos resultados obtidos. Esta etapa deve permitir que um computador possa utilizar o resultado de forma direta ou então deve ser orientada ao usuário, permitindo sua visualização.

As etapas de representação inicial dos objetos (Figura 1-a) e definição da medida de similaridade (Figura 1-b) ainda podem ser modificadas com base nos dados gerados pela etapa: apresentação do resultado (Figura 1-d). Este processo é conhecido como realimentação ou, no inglês, *feedback* e possui o objetivo de especializar o processo para o problema estudado.

2.2. Clusterização Particionada

De acordo com Carlantonio (2001), os algoritmos de clusterização por particionamento dividem a base de dados em *k* grupos, onde o número *k* é fornecido pelo o usuário. Desta forma, o algoritmo de particionamento define os *k* centróides, referencial central dos clusters, com base na média dos *n* objetos dos *k clusters* formados. Os objetos, por sua vez, são divididos entre os *k* clusters de acordo com a medida de similaridade adotada. Sendo assim, cada elemento pertencerá ao cluster que forneça o menor valor de distância entre o objeto analisado e o centróide. A técnica de particionamento escolhida para estudo foi o algoritmo k-means.

2.2.1. K-Means

O algoritmo k-means (também conhecido como: k-médias) é uma técnica de clusterização do grupo particionado. Esta técnica caracteriza-se por: não utilizar um supervisor para definir previamente os padrões que serão gerados; ser dependente de uma entidade externa que informe qual a quantidade *k* de clusters será formada (é daí que vem a primeira letra do nome *k-means*); possui o objetivo de criar *k* grupos (*clusters*, partições) iniciais, e em seguida, utilizar uma técnica de realocação iterativa baseada em similaridade, de forma a melhorar a posição dos centróides dentro de cada grupo. (PIMENTEL, VILMA e OMAR, 2003).

2.2.2. Medidas de Similaridade

Segundo Doni (2004), a medida de similaridade aplicada aos objetos é expressa como uma função, ou métrica, que mede a distância do objeto aos centróides dos *k* clusters gerados. Para calcular essa distância são utilizados os atributos (coordenadas) dos objetos analisados. Estes atributos, por sua vez, podem ser armazenados em diversos tipos de dados, por exemplo: intervalo de escala, binário, categórico, ordinal, ou ainda

uma combinação utilizando esses tipos. Roses (2002), afirma que a distância medida entre pontos de cada objeto equivale ao nível ou grau de similaridade entre eles. Esta medida é utilizada para identificar em qual cluster um objeto deve ser alocado.

Conforme salientado por Carlantonio (2001) e Jain, Murty e Flynn (1999), uma das funções de cálculo de similaridade mais utilizada é conhecida como função de Minkowski, definida pela (equação 1):

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q} \qquad (1)$$

Caso q = 2 essa distância é conhecida como distancia Euclidiana (quadrática) e é definida por: (Equação 2). Essa foi a função adotada para os testes:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \qquad (2)$$

2.2.3. O algoritmo K-means

O algoritmo k-means trabalha da seguinte forma: primeiro, é feita uma seleção aleatória para definir qual elemento será escolhido para assumir o papel de centróide dos *k* clusters (Figura 2-a) ou estes centróides são gerados randomicamente com base nas coordenadas máximas e mínimas dos objetos analisados. Segundo, para cada um dos objetos remanescentes, é calculada similaridade entre os objetos e os centróides, a partir desta medida o objeto é atribuído ao cluster mais similar (Figura 2-b). Terceiro, com base nos objetos dos clusters gerados é computada as novas médias (centróides) para cada cluster (Figura 2-c). Este processo se repete até que a função critério venha a convergir (Figura 2-d) (CARLANTONIO, 2001).

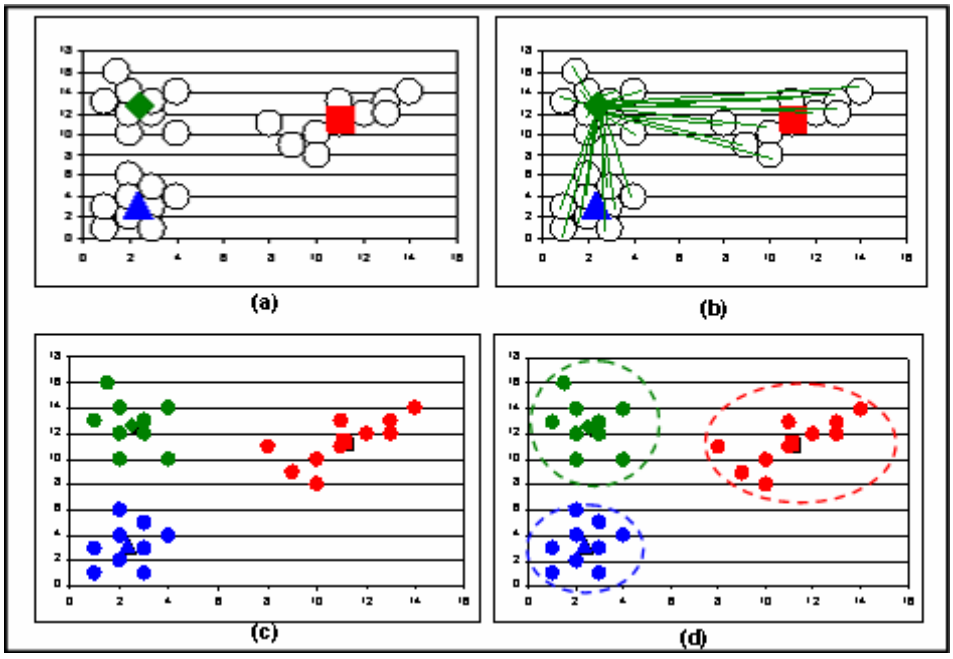

Figura 2: Representação do Algoritmo K-Means

2.2.4. Vantagens e desvantagens do método k-means

O método k-means tem como vantagens (BERKHIN, 2002): Ser relativamente escalável e eficiente para grandes conjuntos de dados. O método freqüentemente termina num local ótimo. Entretanto, este método só pode ser aplicado quando a média (centróide) de um cluster pode ser definido. Isto pode não ser o caso em algumas aplicações, que utilizam dados com atributos categóricos (nominais) estão envolvidos. BERKHIN (2002) comenta que a abordagem por k-means é sensível à partição inicial, gerada pela escolha aleatória dos centróides. A técnica k-means necessita que número *k* de clusters seja informado com antecedência. Além disso, ele é sensível a ruídos, visto que pequeno número de tais dados pode influenciar, substancialmente, o valor médio das coordenadas. O método k-means não é adequado para descobrir clusters com formas não convexas, pelo fato deste método gerar *clusters* com figuras circulares, este problema é conhecido como problema da superposição de classes. Este fenômeno ocorre quando a base de dados analisada possui muitos pontos de confusão (Figura 3). Pontos de confusão são objetos que são difíceis de determinar sua similaridade. Estes pontos são da natureza do problema (da base de dados analisada).

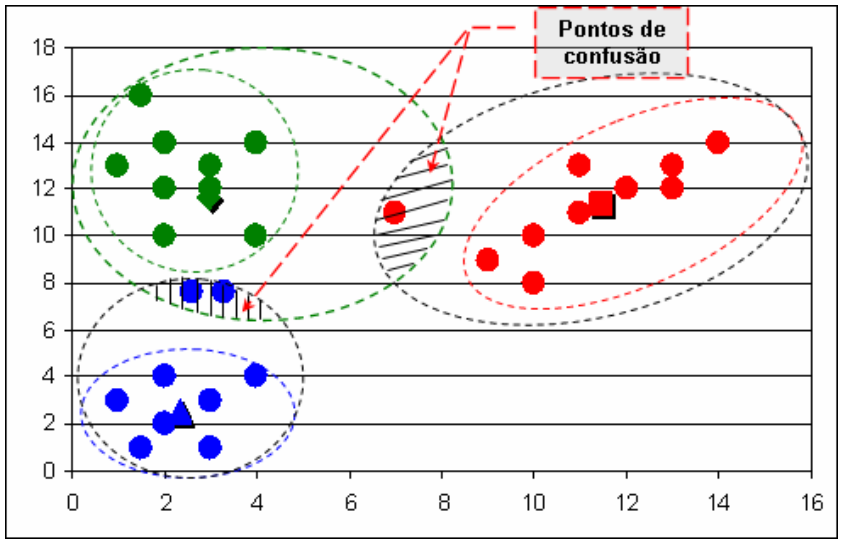

Figura 3: Pontos de confusão

3. Bases de dados: IRIS FLOWER, VOTE e HEART DISEASE

Todas as bases de dados descritas podem ser obtidas no repositório de bases de dados para descoberta de conhecimento da Universidade da Califórnia, Irvine (UCI Machine Learning Repository) (Asuncion e Newman, 2007).

A base de dados IRIS FLOWER consiste em um conjunto de 150 instâncias. Cada instância cinco atributos, sendo quatro do tipo numérico e um do tipo categórico. Esta base de dados é dividida em três classes: Iris-Setosa, Iris-Virginica e Iris-Versicolor, cada uma possuindo 50 instancias. O atributo do tipo categórico não participa do processo de clusterização, entretanto define a qual classe a instância pertence.

A base de dados HEART DISEASE consiste em um conjunto de 270 instâncias. Cada instância possui 13 atributos, sendo cinco do tipo real (idade, pressão do sangue,

soro colesterol, classificação do coração, e velho pico) e os oito restantes são do tipo categórico ou binário. Esta base de dados possui duas classes de dados: ausência e presença de problemas no coração, sendo respectivamente 55.56% e 44.44% o percentual de instâncias que pertencem às classes.

A base de dados VOTE consiste em um conjunto de 435 instancias sendo que cada instância possui 17 atributos, sendo os 16 primeiros atributos do tipo booleano e o 17º atributo, o atributo que define a classe da instância, este pode ser: republicano ou democrata. A base de dados está divida em 267 instâncias da classe republicano e 168 instâncias da classe democrata.

4. Propostas de Melhoria do Processo de Clusterização K-Means

4.1. Primeira Hipótese de Melhoria – Inicialização dos centróides no centro

A hipótese de inicializar os centróides no centro surgiu com a suspeita de que a forma original de inicialização dos centróides aplicada pela técnica k-means, a qual é feita aleatoriamente (Figura 4), tem influência no resultado final do processo de clusterização. Esta hipótese também tem como fator motivacional uma provável redução na quantidade de iterações necessárias para que os centróides estagnem e consequentemente os clusters sejam definidos.

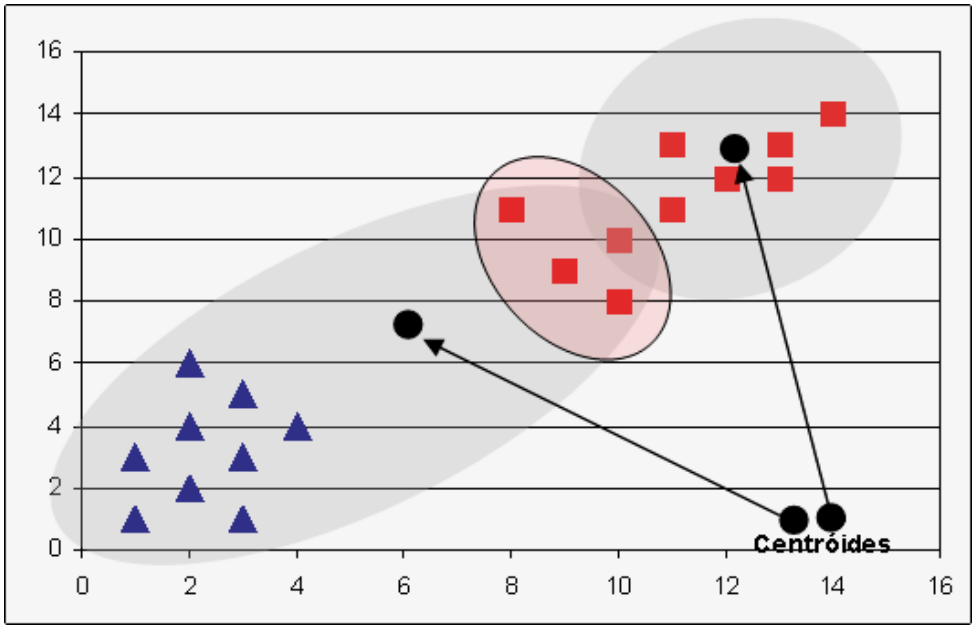

Figura 4: Inicialização Randômica dos Centróides

Portanto, a proposta da primeira hipótese refere-se à seguinte modificação no primeiro passo do algoritmo k-means. O algoritmo varrerá todo o espaço de objetos (pontos ou dados), encontrando em cada coordenada o seu valor máximo e mínimo, e para cada coordenada é feita média entre esses valores, após esta média ser encontrada é aplicada uma função para gerar um número randomicamente. O domínio da função randômica é definido pelo conjunto: R = (0, ..., 2), sendo qualquer que seja o elemento de R dentro do intervalo determinado pertencente ao conjunto dos números reais. Então, o resultado do módulo da média é somado ao resultado do número gerado

randomicamente e este processo é repetido para cada coordenada do ponto até serem formados os k centróides iniciais.

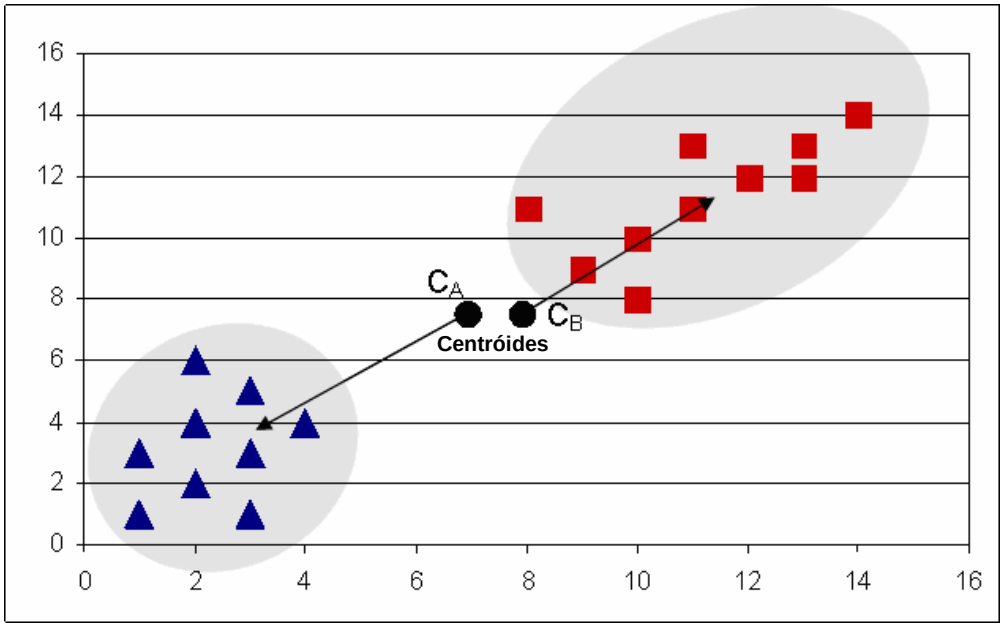

Figura 5: Inicialização no Centro

Supondo que os objetos de A = {(1;1), (2;2), (3;3), (4;4), (3;5), (2;6), (1;3), (3;1), (2;4)} e os objetos de B = {(10;10), (11;11), (12;12), (13;13), (14;14), (9;9), (10;8), (11;13), (13;12), (8;11)}, a aplicação desta hipótese resultaria nos centróides com as posições C_A = (6,947368; 7,473684) e C_B = (7,947368; 7,473684), levando em consideração que a função que gera um número randômico, adicionaram os valores das coordenadas dos centróides os seguintes valores: C_A = (0; 0) e C_B = (1; 0) (Figura 5).

4.2. Segunda Hipótese de Melhoria – Aplicação Automática de Pesos (AAP)

A hipótese da aplicação automática de pesos surgiu da necessidade de criar um deslocamento virtual nos pontos para que estes ficassem mais próximos do centróide mais similar durante as iterações geradas pela técnica k-means.

O objetivo deste deslocamento é possibilitar a criação de figuras circulares, diferentes das figuras geradas pelo k-means, provendo a separação virtual dos objetos de classes distintas. Visando alcançar este objetivo, o presente trabalho desenvolveu a técnica AAP (Figura 6).

Esta técnica é aplicada no k-means após o fim das iterações do mesmo, pois a partir dos clusters formados por esta técnica serão gerados pesos. Estes, que por sua vez, são obtidos através de um estudo modal de cada coordenada dos pontos analisados.

Após os pesos serem definidos, estes são aplicados, em novas iterações k-means, para determinar as novas posições (virtuais) que os pontos assumirão. Conseqüentemente os centróides também mudaram de posição.

Este processo se repete até que os centróides estagnem ou até que o limite de iterações seja alcançado. Estas mudanças acarretarão na geração de novos clusters com características mais similares que as geradas pelo k-means.

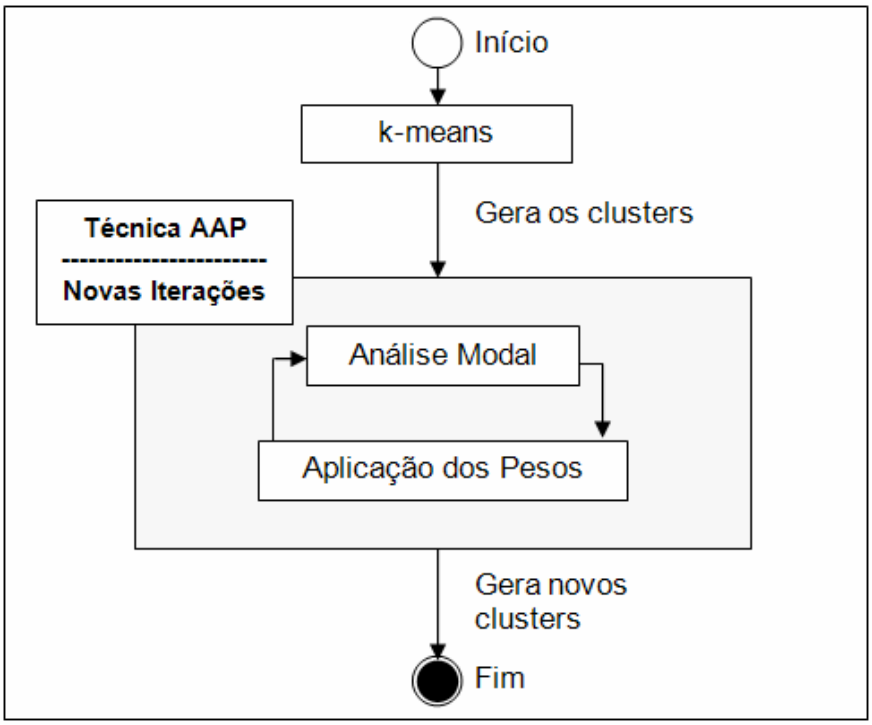

Figura 6: Técnica AAP

4.2.1. Análise modal das coordenadas

A análise modal tornou-se necessária para o processo da AAP, pois nem todas as bases de dados são suficientemente balanceadas para o cálculo da média ou mediana, muitas vezes objetos considerados *outliers* influenciam negativamente no resultado final desses cálculos, portanto para evitar a influência negativa dos *outliers* foi definido que uma medida de posição seria mais apropriada para processo, do que uma medida formal de dispersão, pois sendo a moda o valor que ocorre com maior freqüência em uma série de valores, evitar-se-ia que durante o processo de geração dos pesos fossem incluídos objetos considerados outliers (Figura 7).

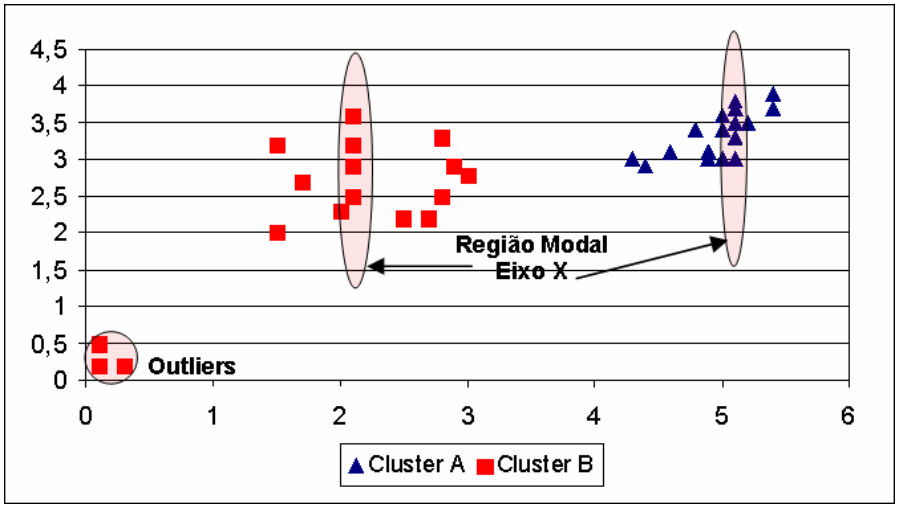

Figura 7: Objetos outliers e região modal

Supondo o Cluster A e o Cluster B (Figura 7) (Tabela 1), verifica-se que a moda de cada coordenada é respectivamente, MoCAx = (5,1), MoCAy = (3), MoCBx = (2,1) e MoCBy = (2,9).

Tabela 1: Região Modal / Pesos

	CLUSTER			
	A		B	
	x	y	x	y
AMOSTRAS	4,3	3	0,1	0,2
	4,4	2,9	0,1	0,5
	4,6	3,1	0,3	0,2
	4,8	3,4	1,5	3,2
	4,9	3	1,5	2
	4,9	3,1	1,7	2,7
	5	3,6	2	2,3
	5	3,4	2,1	2,9
	5	3	2,1	3,2
	5	3,4	2,1	2,5
	5,1	3,5	2,1	2,9
	5,1	3,5	2,1	3,6
	5,1	3,8	2,5	2,2
	5,1	3,7	2,7	2,2
	5,1	3,3	2,8	3,3
	5,2	3,5	2,8	2,5
	5,4	3,9	2,9	2,9
	5,4	3,7	3	2,8
MODA (PESOS)	**5,1**	**3**	**2,1**	**2,9**

Às vezes, é possível distinguir claramente “picos” na freqüência dos valores registrados. Casos como estes onde dois ou mais valores ocorrem com a mesma freqüência máxima são conhecidos como conjunto multimodal, para esses casos deve-se levar em consideração todas as localizações e fazer a média entre elas para que o valor do peso seja aplicado. Desta forma, verifica-se que o processo de análise modal consiste em verificar os elementos que se repetem com maior freqüência em cada coordenada, armazenando-os em um vetor que será posteriormente utilizado como peso durante as novas iterações do k-means.

4.2.2. Proposta da AAP

Com o termino do processo da análise da modal, o vetor W_i contendo os valores dos pesos já estará gerado, possibilitando que estes pesos sejam aplicados no cálculo da medida de similaridade. Entretanto o objetivo desta proposta é criar um deslocamento virtual dos pontos, aproximando-os do centróide mais similar a eles. Para isto é necessário criar um peso que seja inversamente proporcional (equação 3) aos valores modais encontrados. Desta forma, será criada uma distorção virtual no espaço vetorial convexo gerado pelo processo de clusterização k-means.

$$d(i,j) = \sqrt[2]{\left(\frac{1}{w_1}\left|x_{i1} - x_{j1}\right|\right)^2 + \left(\frac{1}{w_2}\left|x_{i2} - x_{j2}\right|\right)^2 + \ldots + \left(\frac{1}{w_p}\left|x_{ip} - x_{jp}\right|\right)^2} \qquad (3)$$

Este processo se repete até que o limite de iterações seja alcançado ou até que o centróide estagne. Desta forma, uma nova análise modal é gerada a cada iteração proporcionando uma maior separabilidade entre os objetos de classes distintas. Esta proposta adiciona ao processo de clusterização mais duas fases, sendo elas: a análise modal e a aplicação dos pesos.

5. Testes e Resultados

5.1. Equipamento e linguagem utilizados nos testes

Foi utilizado para execução dos testes um notebook ACER da série Aspire 5612NWLMi com a seguinte configuração: Processador Intel Core Duo T2300 'Centrino' (2x 1.66GHz 2MB L2 FSB 667), Gravador de DVD Super Multi Dual layer, Rede wireless 802.11 b/g, rede ethernet 10/100, Modem 56Kbps V.92, Disco rígido de 120GB, Memória RAM 1GB DDR2 (2X 512 DDR2 533).

PHP foi à linguagem de programação escolhida para desenvolvimento da aplicação. Esta, por sua vez, é uma linguagem de programação, estruturada e orientada a objetos, de domínio específico, ou seja, seu escopo se estende a um campo de atuação que é o desenvolvimento web, embora tenha a variante PHP-GTK, a qual é utilizada no desenvolvimento de aplicações desktop, o propósito principal desta linguagem é implementar soluções web velozes, simples e eficientes. O tempo médio de carregamento, da página responsável pela clusterização dos objetos, foi de 0,4932 segundos variando de 0,03 segundos a 0,99 segundos. Este resultado foi concebido após serem realizados 100 testes de carregamento.

5.2. Características dos experimentos

Os experimentos foram divididos em três etapas. Nestas etapas foi analisado, nas bases IRIS FLOWER, VOTE e HEART DISEASE, o resultado da clusterização executada pela técnica k-means (Etapa 1), bem como as propostas de melhoria apresentas neste trabalho (Etapa 2 e Etapa 3), durante as etapas 2 e 3, foram feitos os confrontos dos resultados obtidos pela técnica k-means e os resultados obtidos através das propostas de melhoria. Os experimentos mostraram que a proposta AAP apresentada neste trabalho surtiu efeito positivo no resultado final do processo de clusterização (Tabela 2), comprovando que as hipóteses apresentadas procedem.

Tabela 2: Resultado dos experimentos

EXPERIMENTO	BASES					
	IRIS FLOWER		VOTE		HEART DISEASE	
	MÉDIA		MÉDIA		MÉDIA	
	ABRANGÊNCIA	ACURÁCIA	ABRANGÊNCIA	ACURÁCIA	ABRANGÊNCIA	ACURÁCIA
K-MEANS	88,91%	89,79%	89,45%	86,27%	57,75%	58,06%
CENTRO	89,33%	90,72%	89,30%	87,36%	58,00%	58,40%
PESO	**91,33%**	**92,59%**	**96,79%**	**95,72%**	**61,50%**	**61,40%**

Entretanto, não houve diferença plausível na quantidade de iterações necessárias para que os centróides dos clusters estagnassem (Tabela 3), pois a diferença entre um experimento e o outro foi de no máximo duas iterações e o tempo médio de carregamento, da página responsável pela clusterização dos dados, foi de 0,4932 segundos. Este fato nega uma das suspeitas da proposta de Inicialização no Centro, a qual informa que esta quantidade de iterações reduziria, caso os centróides fossem inicializados no centro.

Tabela 3: Resultado dos experimentos

EXPERIMENTO	BASES					
	IRIS FLOWER		VOTE		HEART DISEASE	
	VARIAÇÂO DAS ITERAÇÔES		VARIAÇÂO DAS ITERAÇÔES		VARIAÇÂO DAS ITERAÇÔES	
	MIN	MÁX	MIN	MÁX	MIN	MÁX
K-MEANS	3	14	1	12	4	15
CENTRO	2	16	2	11	5	15

6. Conclusão

O problema estudado neste trabalho refere-se à tentativa de melhoria no processo de clusterização da técnica k-means, a qual se caracteriza principalmente por: não utilizar um supervisor para definir previamente os padrões que serão gerados; ser dependente de uma entidade externa que informe qual a quantidade k de clusters será formada.

Este estudo teve como desafio a comprovação das propostas mencionadas neste trabalho, sendo elas: a Inicialização no Centro e a Aplicação Automática dos Pesos. A primeira se subdivide em duas hipóteses, a hipótese que a inicialização dos centróides no centro resultaria na convergência de uma melhor solução de clusterização, esta sendo comprovada neste trabalho e a hipótese referente à quantidade de interações necessárias para que uma solução fosse encontrada, esta última hipótese foi negada, pois de acordo com os testes realizados a diferença encontrada entre a quantidade de interações do k-means normal e a quantidade de iterações do k-means utilizando a inicialização no centro foi pequena. A segunda proposta, a qual se refere ao método da Aplicação Automática dos Pesos, teve resultado satisfatório (Tabela 2).

As vantagens apresentadas pelas propostas podem ser facilmente percebidas analisando apenas os resultados alcançados, haja vista, caso um sistema de detecção problemas no coração utilizasse como base a técnica k-means, para encontrar padrões em uma população de pessoas, a melhoria apresentada neste trabalho determinaria com maior precisão se a pessoa analisada enquadrar-se-ia na região de pessoas com problemas no coração ou de pessoas saudáveis.

Entretanto, as desvantagens encontradas para cada proposta são: O fato da inicialização acontecer sempre no centro pode fazer com que a solução estagne em um pico local, em quanto à solução randômica pode alcançar outros picos. Entretanto, os testes executados nas bases de dados informaram uma melhora no processo de clusterização. Em relação à segunda proposta (AAP) a desvantagem encontra-se na

quantidade de iterações necessárias para que a solução seja encontrada, haja vista, o processo de aplicação dos pesos começa somente quando o a clusterização feita pelo k-means termina.

Como proposta futura pretende-se fazer outros tipos de testes com a estratégia dos pesos, com o seguinte diferencial, ou invés de aplicar os pesos em todas as coordenadas é pretendido aplicar os pesos nas coordenadas mais importantes, o problema encontra-se em como definir qual ou quais coordenadas são importantes.

7. Referencias

Asuncion, A. e 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.

Berkhin, P. (2002), "Survey of Clustering Data Mining Techniques", Accrue Software, Inc.

Carlantonio, L. M. Di (2001). "Novas metodologias para clusterização de dados". Dissertação (Mestrado)-Programa de Pós-Graduação em Engenharia, Universidade Federal do Rio de Janeiro, RJ, Brasil.

Doni, M. V. (2004). "Análise de Cluster: Métodos Hierárquicos e de Particionamento", Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie, São Paulo, SP, Brasil.

Jain, A.K., Murty, M.N., Flynn, P.J. (1999). "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, Setembro.

Ochi L. S., Dias C. R., Soares S. S. F. (2004). "Clusterização em Mineração de Dados", Instituto de Computação - Universidade Federal Fluminense (IC - UFF), Niterói, Rio de Janeiro, Brasil.

Pimentel, E. P., França, V. F. De, Omar, N. (2003). "A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização", Instituto Tecnológico da Aeronáutica (ITA), São José dos Campos, SP, Brasil.

Roses, C. F. (2002). "Um estudo das condições sócio-econômicas de municípios gaúchos através da análise de cluster", Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, RS, Brasil.