APLICAÇÃO DE *FUZZY CLUSTERING* A BANCO DE DADOS DE AMOSTRA DOMICILIAR DA POPULAÇÃO DA ILHA DO GOVERNADOR, R.J.

J. M. C. DA ROSA[1], R. TANSCHEIT[1], M. M. VELLASCO[1], A. ZANINI[1], C. H. KLEIN[2], K. V. BLOCH[3], A. R. NOGUEIRA[4], L. H. SALIS[4], N. A. SOUZA E SILVA[5]

[1]*DEE/PUC-Rio, CP 38.063, 22.452-970 Rio de Janeiro, RJ*
[2]*Escola Nacional de Saúde Púbica (ENSP/FIOCRUZ)*
[3]*Departamento de Medicina Preventiva, Faculdade de Medicina/UFRJ*
[4]*Hospital Universitário Clementino Fraga Filho/UFRJ*
[5]*Departamento de Clínica Médica, Faculdade de Medicina/UFRJ*
E-mails: [marley,ricardo]@ele.puc-rio.br

Resumo – No presente estudo utiliza-se a técnica de mineração de dados para analisar o banco de dados obtido do "Estudo Multicêntrico sobre Hipertensão Arterial e outros fatores de risco cardiovascular na população da Ilha do Governador, Rio de Janeiro". O banco de dados original foi obtido através de questionário e medidas objetivas representando: informações demográficas, características constitucionais e sócio-econômicas, hábitos de vida, controle da pressão arterial e consumo de medicamentos. Foram entrevistados 1270 casos após amostragem domiciliar randômica de conglomerados em dois estágios, estratificada por classe sócio-econômica. No presente estudo são selecionados apenas nove variáveis para realizar esta mineração dos dados. Utiliza-se a técnica de *fuzzy clustering*, cuja aplicação permite a obtenção de grupos de indivíduos que apresentam comportamento semelhante em um conjunto de atributos. Com a utilização de informações de seis atributos, os resultados indicam a formação de três *clusters* homogêneos internamente e significativamente diferentes entre si, conforme testes estatísticos. A separação de indivíduos de uma população em *clusters* pode ser utilizada, por exemplo, para identificar possíveis relações causais em doenças ou para separar grupos de indivíduos que possam sofrer intervenções de saúde diversas. As variáveis que identificam o *cluster* podem indicar também o uso de formas ou métodos diferentes de intervenção.

Abstract – This works consists of an application of data mining to the analysis of a database originated from the "Multicentric Study on Hypertension and other cardiovascular risk factors in the population of Ilha do Governador, Rio de Janeiro". The original database was generated through questionnaires and objective measures representing demographic information, socio-economical characteristics, life habits, arterial pressure control and medication intake. Interviews considered 1270 cases, after random sampling in socio-economical classes. The present study takes into account nine variables for data mining. The technique chosen is fuzzy clustering, which gives as a result groups of individuals with similar behaviour regarding a given set of attributes. By using six attributes, three homogeneous clusters – and significantly different from each other – are formed. The separation of individuals in clusters may be used, for example, to identify possible causal relations among diseases or to tell apart groups of individuals that may be liable to medical intervention. The variables that identify the cluster may indicate forms or methods of intervention.

Keywords – data mining; fuzzy clustering; hypertension

1 Introdução

As Doenças Cardiovasculares (DCV) são a principal causa de morte no país. No Rio de Janeiro em 1998 as DCV representaram 30% do total de óbitos, o triplo do percentual de óbitos ocasionados pela segunda causa (câncer). Entre as DCV, as que mais matam são o Acidente Vascular Cerebral (AVC) e a Doença Cardíaca Isquêmica (DCI). Estas doenças são, também, responsáveis pelo maior número de aposentadorias por doença, são a terceira causa de internações (12% do total), e representam o maior gasto com estas internações (17% do total). Conhecem-se inúmeros fatores – denominados fatores de risco – que aumentam a probabilidade de ocorrência da DCI e do AVC. Entre estes fatores podem ser citados como principais: a Hipertensão Arterial (HA), o Diabetes Mellitus (DM), o Tabagismo, o uso excessivo de bebidas alcoólicas, o sedentarismo, as dislipidemias, a obesidade, entre outros. Dentre os fatores de risco cardiovascular a HA assume particular importância por ter alta prevalência, estar associada a 85% dos casos de AVC e a 60% dos casos de infarto do miocárdio e, entre as causas de aposentadoria por doença, por ser a principal, com 19% do total de casos aposentados. Em estudo realizado na Ilha do Governador (I.G.), no Rio de Janeiro (Klein, 1995b) em amostra domiciliar randômica estratificada por classe sócio-econômica encontramos uma prevalência de HA de 38%. Esta prevalência aumenta com a faixa etária, principalmente entre as mulheres, atingindo a mais de 50% na faixa acima de 60 anos de idade, atestando a importância da HA como um dos principais problemas de saúde pública no País. Sabe-se também que a prevalência de HA varia bastante se forem comparadas diversas populações ou grupos sociais distintos. Assim, entre os índios Yanomami, não se encontra nenhum hipertenso (Mancilha-Carvalho, 1991) e entre os índios Terenas a prevalência de HA encontrada foi de 7% (Mancilha-Carvalho, 1983)

A Figura 1 mostra a nítida diferença entre as curvas de distribuição da pressão arterial (PA) em populações distintas. Estas diferenças indicam a necessidade de buscar os fatores envolvidos na elevação da pressão arterial e as características de sua aglomeração em grupos populacionais ou etnias

diversas para que se possam adotar métodos diversos para controlá-los ou eliminá-los, de acordo com as características identificadas, reduzindo a prevalência da HA e suas conseqüências.

A pressão arterial é controlada por diversos sistemas fisiológicos, tanto vasodilatadores quanto vasoconstrictores, e que mantêm a pressão arterial dentro de certos limites. Estes sistemas fisiológicos estão sob controle genético e sofrem a influência de fatores ambientais. Possuem, quando estimulados, tempo de resposta variável, seja imediata, seja mais tardiamente. Conhecem-se alguns fatores de risco para a hipertensão arterial, ou seja, fatores que aumentam a probabilidade da pressão elevar-se com o passar do tempo ou que desregulam os mecanismos de homeostase da PA. Dentre estes fatores podem ser mencionados: a idade (com exceção dos Yanomami, onde esta elevação com a idade não ocorre), a obesidade, o DM, as dislipidemias, a ingestão excessiva de sal, o uso abusivo, diário, de bebidas alcoólicas e condições de vida determinadas por baixo nível sócio-econômico incluindo o desemprego e o analfabetismo. O fumo, embora não seja uma causa para a elevação da pressão arterial, aumenta consideravelmente os seus riscos quando a ela associada (Hart, 2000). Portanto, a elevação da pressão arterial é determinada por uma rede complexa de causalidade ou por uma interação complexa entre genes, organização social e meio ambiente. Desta complexa interação pode resultar a elevação, intermitente ou constante, da pressão arterial, e desta elevação resultam problemas clínicos graves como o AVC, a DCI, a Insuficiência cardíaca, a Insuficiência renal que acabam ocasionando a morte do indivíduo. O controle da HA reduz a probabilidade destes eventos ocorrerem.

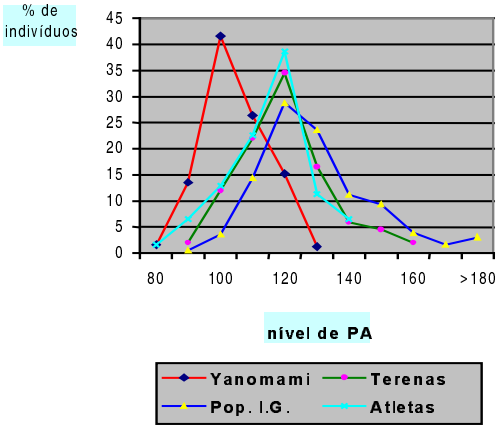

Figura 1. Distribuição de pressão arterial sistólica no sexo masculino em diferentes populações

O presente trabalho tem como objetivo extrair informações do banco de dados obtido originado da pesquisa sobre hipertensão arterial na população da Ilha do Governador, realizada pelo Ministério da Saúde com apoio do CNPq e executada pela Universidade Federal do Rio de Janeiro (Faculdade de Medicina e Hospital Universitário Clementino Fraga Filho) e pela Escola Nacional de Saúde Pública (FIOCRUZ), em 1990-1992. Para isto será utilizada a técnica de *fuzzy clustering.*

2 O Banco de Dados

O banco de dados original foi construído no inquérito domiciliar da população da Ilha do Governador (Klein, 1995a). Aquele estudo foi idealizado com o objetivo de estimar a prevalência de HA na população adulta (acima de 20 anos de idade) e para analisar as possíveis associações da pressão arterial com algumas variáveis pré-definidas. Foram selecionados e entrevistados um total de 1270 indivíduos, moradores de 750 domicílios da I.G., nos quais aplicou-se questionário padronizado e realizaram-se medidas objetivas. Para o presente estudo, foram escolhidas inicialmente nove variáveis do banco de dados original: idade, sexo, peso corporal, altura, pressão arterial sistólica e diastólica, renda familiar, escolaridade, consumo de cigarros.

Na Tabela 1 são apresentadas estimativas para as principais estatísticas de sete das variáveis selecionadas do banco de dados original.

Tabela 1 – Estatísticas Descritivas da Amostra da População de Adultos (acima de 20 anos) da I.G.

Variável	Mediana	Média	Mínimo	Máximo	Desvio padrão
Idade	41,00	43,14	20,00	91,00	15,44
Peso	65,00	66,41	32,00	134,00	13,36
Altura	162,00	162,59	136,00	198,00	9,84
Pressão Sistólica	128,00	131,22	82,00	254,00	21,53
Pressão Diastólica	80,00	81,82	50,00	152,00	11,71
Consumo Cigarros	2,00	9,61	0,00	88,00	14,2
Renda	132,88	289,22	0,00	3653,93	437,05

As características medianas de um indivíduo amostrado são: 41 anos de idade, 65 kg de peso, 1,62 m de altura, pressão arterial sistólica de 128mmHg, pressão arterial diastólica de 80 mmHg, renda de 132,88 dólares e consumo de cigarros igual a 2.

Em relação a outros dados básicos dos entrevistados, observa-se na Tabela 2 que 55,6% da amostra é composta por mulheres e mais da metade dos indivíduos sequer concluiu o segundo grau. A prevalência de hipertensos (PA ≥ 160/95 mm Hg) estimada para esta população foi de 25%. Considerando, no entanto, como critério de HA o valor da PA≥140/90 mm Hg, ter-se-ia uma prevalência de hipertensão arterial de 38% nesta população.

Tabela 2 – Distribuição dos Indivíduos segundo Sexo, Escolaridade e Prevalência de Hipertensos

Variáveis	Categorias	Freqüência Absoluta	Freqüência Relativa
Sexo	Masculino	547	44,44 %
	Feminino	684	55,56%
Escolaridade	Analfabeto	41	3,33%
	Auto-Aprendizado	26	2,11%
	Primeiro Grau Incompleto	384	31,19%
	Primeiro Grau completo	216	17,55%
	Segundo Grau Completo	341	27,70%
	Terceiro Grau Completo	223	18,12%
Hipertensão	Não Hipertensos	926	75,22%
	Hipertensos	305	24,78%

Salienta-se que uma possível associação ou relação causal da variável de interesse – hipertensão arterial – com as variáveis contidas no banco de dados não pode ser identificada com uma simples descrição global. Em (Bloch, 1994), por exemplo, analisam-se os dados através de um modelo de regressão logística que aponta uma forte associação entre hipertensão e obesidade em indivíduos do sexo masculino e nos jovens. No presente trabalho, de modo alternativo, busca-se encontrar associações entre as variáveis através da formação de agrupamentos, ou *clusters,* de indivíduos que apresentem comportamento semelhante em relação a determinadas características mensuradas.

3 Algoritmo de *Clusterização Fuzzy*

A análise de *cluster* é uma técnica exploratória de dados que tem por objetivo formar agrupamentos de objetos semelhantes em um banco de dados. O conceito de *clusterização* difere do conceito de classificação no sentido de que a análise de *cluster* é uma técnica mais "primitiva", na qual nenhuma suposição é feita a respeito dos grupos, assim como o seu número e estrutura (Jonhson, 1998). Os *clusters* são obtidos por intermédio da aplicação dos conceitos de similaridade e de distância.

Ao se trabalhar com valores na reta dos números reais, pode-se intuir que a distância entre dois números quaisquer será mensurada através da diferença em módulo. No contexto deste trabalho, no entanto, empregam-se vetores no espaço p-dimensional e, assim, o conceito de distância entre dois objetos depende de uma série de fatores, entre eles a natureza da variável (discreta, contínua, binária) e as escalas de mensuração (ordinal, intervalar). No presente trabalho a medida de distância adotada é a distância euclidiana, caracterizada na equação abaixo, que define a distância entre dois vetores p-dimensionais.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$

Ao contrário dos métodos usuais de *clusterização* de dados, na c*lusterização fuzzy* o indivíduo não pertence somente a um grupo, mas pode pertencer a todos os grupos com diferentes graus de pertinência (Klir, 1988). A restrição adotada nesta metodologia é que a soma dos graus de pertinência de um indivíduo aos grupos seja igual a 1. Tal restrição vai ao encontro da própria definição de *cluster*: agrupamento de indivíduos similares segundo algumas características. Portanto, seria incoerente a um indivíduo pertencer a dois grupos com graus de pertinência muito altos (fazendo com que a soma ultrapasse a unidade), o que significaria altos graus de similaridade concomitantes. O algoritmo utilizado neste trabalho, denominado *fuzzy c-means*, foi proposto por Bezdek (Bezdek, 1984).

No processo de *clusterização* cada um dos *n* indivíduos tem associado a si um vetor $\mathbf{x_j}$ contendo os atributos. A técnica de *fuzzy clustering* é empregada para que os *n* vetores $\mathbf{x_j}$ sejam agrupados em *c clusters*. O algoritmo para obtenção dos *clusters* minimiza a seguinte função objetivo:

$$J(u_{ij}, v_k) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| \mathbf{x_j} - v_i \right\|^2 \qquad (1)$$

onde $m \in [1,\infty)$ é o coeficiente fuzzy responsável pelo grau de *fuzzificação* dos elementos de $\mathbf{x_j}$ e v_k é o centróide do k-ésimo *cluster*.

O resultado da *clusterização fuzzy*, portanto, pode ser expresso através da matriz $U = [u_{ij}]$, para i=1, ..., c e j=1, ..., n; onde u_{ij} é um valor entre 0 e 1 que indica o grau de pertinência de cada elemento x_j em um determinado *cluster i*. Quanto maior for o coeficiente *m*, mais *fuzzy* se torna a matriz *U*. Quando *m* é igual a 1, a função objetivo $J(u_{ij}, v_k)$ é reduzida ao caso *crisp*, que corresponde ao algoritmo de *clusterização k-means.*

De forma analítica, podemos avaliar o mínimo da função objetivo (1), minimizando-a em relação aos argumentos. Assim, é necessária a resolução das equações:

$$\frac{\partial J(u_{ij}, v_k)}{\partial u_{ij}} = 0 \quad e \quad \frac{\partial J(u_{ij}, v_k)}{\partial v_k} = 0$$

Os valores que minimizam a função objetivo são obtidos a partir das seguintes equações:

$$u^{ij} = \frac{\left(1 / \| x^j - v^i \| \right)^{\frac{1}{m-1}}}{\sum_{k=}^{c} \left(1 / \| x^j - v^k \| \right)^{\frac{1}{m-1}}} \qquad (2)$$

$$v^i = \frac{1}{\sum_{j=1}^{n} \left(u^{ij}\right)^m} \sum_{j=1}^{n} \left(u^{ij}\right)^m x^j \qquad (3)$$

O algoritmo propriamente dito pode ser descrito através dos passos seguintes:

1. Inicializar: c (número de *clusters*) , ε (critério de parada), m (coeficiente fuzzy), U^0 (matriz inicial com os graus de pertinência);
2. Calcular os centróides dos *clusters*.
3. Atualizar a matriz U^{k+1}
4. Calcular $\nabla = \| U^{k+1} - U^k \|$. Se $\nabla < \varepsilon \Rightarrow$ fim do algoritmo, se não, $k = k+1$ e voltar ao passo 2.

Para avaliar o poder de discriminação do algoritmo FCM, é utilizada a medida de entropia de Kullback-Liebler (Cover, 1991):

$$KL_j = \sum_{i=1}^{c} p_i \left(\log(p_i / u_{ij}) \right) \qquad (4)$$

Para cada indivíduo j mede-se a distância entre a distribuição dos graus de pertinência aos *clusters* e a situação extrema em que o indivíduo tem igual grau de pertinência aos *clusters* ($p_i = 1/c$), não havendo discriminação.

Como forma global de avaliar a capacidade de discriminação do algoritmo, é calculado o valor médio de *KL* para diversos valores do coeficiente fuzzy m:

$$\overline{KL} = \frac{\sum_{j=1}^{n} KL_j}{n} \qquad (5)$$

Observe-se que, quanto mais próximo de zero estiver o valor de *KL*, menor será a capacidade do algoritmo de discriminar o indivíduo em um *cluster*.

4 Discussão e Resultados

Neste estudo foram considerados 1270 indivíduos, associando-se a cada um deles um vetor contendo os seguintes atributos: x_1 – idade; x_2 – consumo de cigarros industriais; x_3 – peso; x_4 – altura; x_5 – pressão arterial sistólica; x_6 – pressão arterial diastólica. A variável "renda" não foi incluída pois se observou que não possibilitava a discriminação de grupos diferentes quanto à hipertensão – apenas causava o agrupamento de indíviduos segundo sua renda.

Dado que a variável "consumo de cigarros" apresenta vários valores iguais a zero e também devido às diferentes escalas de mensuração das variáveis envolvidas na análise, foi necessário efetuar, de início, uma padronização das variáveis para que todas apresentassem média 0 e desvio padrão igual a 1.

Com diversas tentativas de valores para o coeficiente fuzzy m, verificou-se que valores altos produziam pouca discriminação dos indivíduos nos *clusters*, independentemente do número destes. A Figura 2 apresenta o valor da estatística de Kullback Liebler para m variando de 2 a 20, considerando-se o número de *clusters* igual a 3. Em função da queda muito acentuada, observada após o valor m = 2, este foi escolhido como sendo o "ótimo".

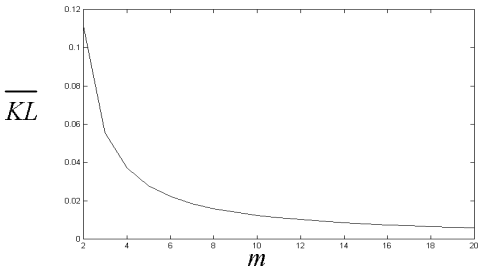

Figura 2 - Média da estatística *KL* segundo o coeficiente m

Através da análise de seis características orgânicas dos indivíduos, da utilização de um coeficiente m = 2 e da pré-fixação do número de *clusters* (c = 3), o resultado obtido apontou três grupos cujas principais estatísticas estão dispostas na Tabela 3**,** que representa o protótipo de cada agrupamento. Os *clusters* foram rotulados com os números 1, 2 e 3; os números entre parênteses indicam o número de indivíduos em cada um.

Tabela 3. Estatísticas descritivas das principais variáveis nos *clusters* obtidos através do algoritmo FCM (m=2 e c=3)

	Idade	Cigarros	Peso	Altura	Pressão Sist.	Pressão Diast.
Cluster 1 (391)	40,40	16,07	77,56	171,94	131,54	84,25
Cluster 2 (478)	33,83	6,29	58,16	159,80	115,62	74,13
Cluster 3 (362)	58,38	7,02	65,25	156,16	151,48	89,37
Média Global	43,14	9,61	66,41	162,59	131,22	81,82

Como o indivíduo pode pertencer a mais de um grupo com diferentes graus de pertinência, para se chegar aos resultados expressos fez-se a associação do indivíduo a um grupo específico pela observação do maior grau de pertinência.

Pode-se observar que os grupos resultantes apresentam características bem definidas. No *cluster* 3, por exemplo, tem-se indivíduos de maior idade e com maior pressão sistólica e diastólica. Este grupo é o segundo em consumo de cigarros e índice de massa corpórea (peso/altura). Já no *cluster* 1, estão indivíduos mais jovens que os do *cluster* 3, mas que fumam mais e têm a segunda maior medida de pressão. São também indivíduos com maior índice de massa corpórea. Por fim, encontram-se no *cluster* 2 indivíduos mais jovens, que consomem menos cigarros, tem menor índice de massa corpórea e, portanto, associados a menores medidas de pressão.

A distribuição de hipertensos nos *clusters* segundo o sexo pode ser observada na Figura 3. Na pesquisa original, foram caracterizados como hipertensos indivíduos com pressão sistólica acima de 140 mmHg e pressão diastólica acima de 90 mmHg. Observa-se que a prevalência de hipertensão arterial é maior no *cluster 3*, enquanto o *cluster 2* apresenta a menor prevalência de hipertensos. Vê-se ainda que, no *cluster* 3, há uma maior prevalência de hipertensos entre as mulheres.

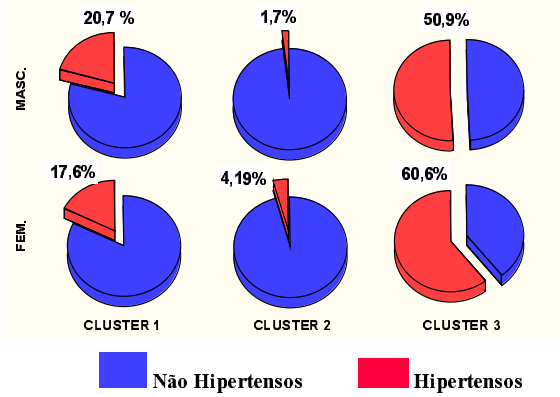

Figura 3 - Hipertensão nos *clusters* segundo o sexo

Na Figura 4 observa-se que o *cluster* 3 é aquele cujos indivíduos apresentam o menor grau de instrução, ao passo que o *cluster* 1 apresenta o maior número de indivíduos com formação universitária.

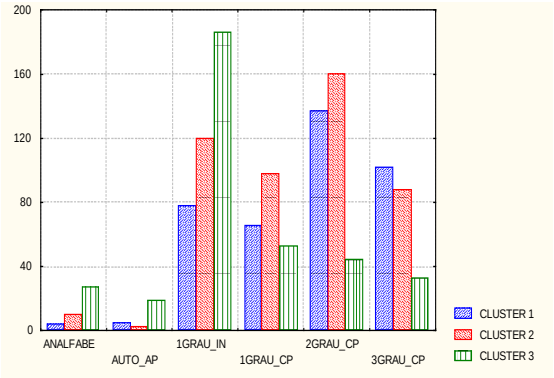

Figura 4 - Escolaridade nos *clusters*

Procurou-se também verificar a distribuição dos graus de pertinência em cada agrupamento, de forma que fosse possível a identificação de possíveis *outliers*. Observando as Figuras 5, 6 e 7, pode-se ver claramente a presença de pontos discrepantes. A partir desta constatação, adotou-se o procedimento de identificar e retirar os *outliers* de cada *cluster*, aplicando-se novamente o *fuzzy c-means* a fim de se verificar a possível ocorrência de mudanças significativas nos agrupamentos.

São apresentadas na Tabela 4 as médias dos valores de cada variável para os cinco indivíduos com os maiores graus de pertinência ao *cluster*, assim como para os cinco com os menores graus.

Identitificados os pontos discrepantes e retirando-os da amostra, verificou-se que não ocorreram mudanças signficativas nas caracaterísticas já comentadas para os *clusters*.

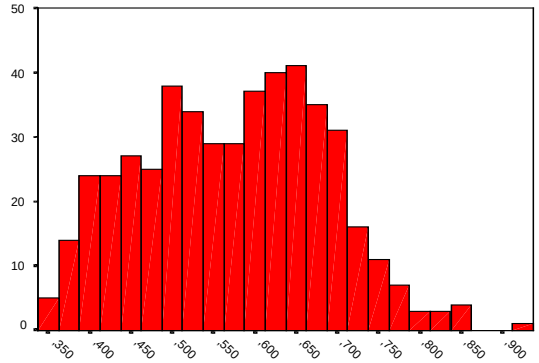

Figura 5 . Distribuição dos graus de pertinência dos indivíduos pertencentes ao *cluster* 1

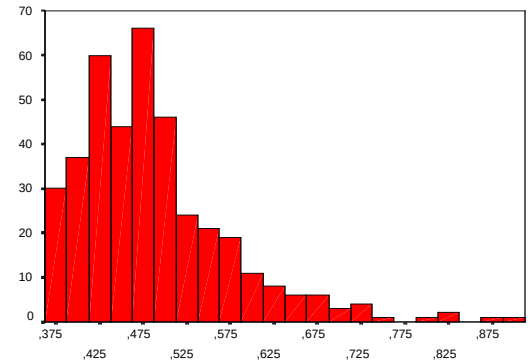

Figura 6 - Distribuição dos graus de pertinência dos indivíduos pertencentes ao *cluster* 2

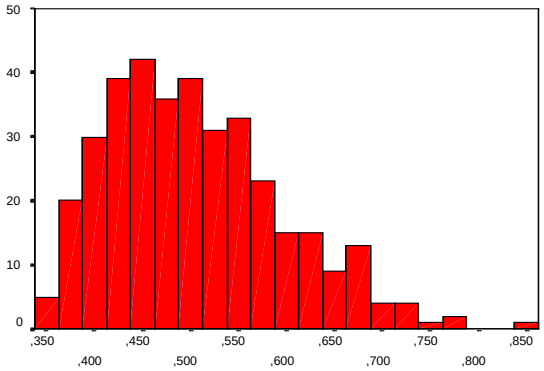

Figura 7 - Distribuição dos graus de pertinência dos indivíduos pertencentes ao *cluster 3*

Tabela 4. Perfil dos indivíduos com graus de pertinência extremos nos *clusters*

	Idade	Cigarros	Peso	Altura	Pressão Sist.	Pressão Diast.
Menores Graus						
Cluster 1	47,00	30,00	63,75	167,50	132,25	79,50
Cluster 2	44,75	10,00	70,00	158,75	123,00	80,50
Cluster 3	46,66	25,00	64,33	154,66	127,33	80,66
Maiores Graus						
Cluster 1	43,00	15,00	72,25	169,75	130,50	83,50
Cluster 2	33,00	4,40	61,40	161,40	117,20	75,60
Cluster 3	57,75	4,50	65,00	158,75	148,00	89,00

Por fim, fez-se também um teste ANOVA (Johnson, 1998) para cada variável, visando identificar diferenças significativas entre as médias dos grupos. Este fato foi constatado a um nível de confiança de 95%.

Com o objetivo de identificar os fatores que proporcionaram a discriminação dos indivíduos nos *clusters*, foi mensurado, através do coeficiente de correlação linear de Pearson (Johnson, 1998), o grau de associação entre os graus de pertinência aos *clusters* e as variáveis utilizadas para a formação dos *clusters.* Observa-se na Tabela 5 que os graus de pertinência ao *cluster 1* estão mais associados às variáveis peso e altura e, conseqüentemente, com características do sexo masculino. Em relação ao *cluster 2,* verifica-se que os graus de pertinência estão associados negativamente com as variáveis idade, pressão sistólica e pressão diastólica, o que caracteriza sujeitos não hipertensos. O *cluster 3* contém indivíduos cujos graus de pertinência estão associados com idade avançada e alta pressão sistólica.

Tabela 5. Correlação linear dos graus de pertinência aos *clusters* com as variáveis utilizadas em sua formação

	Idade	Cigarros	Peso	Altura	Pressão Sist.	Pressão Diast.
Cluster 1	-0,05	0,33	**0,61**	**0,67**	0,11	0,23
Cluster 2	-0,55	-0,21	-0,50	-0,17	**-0,62**	**-0,56**
Cluster 3	**0,71**	-0,04	0,06	-0,38	**0,65**	0,48

5 Conclusões

O algoritmo FCM, aplicado ao vetor de variáveis relacionadas a características dos indivíduos, possibilitou a formação de 3 *clusters* significativamente diferentes entre si. Nestes *clusters,* os protótipos apresentaram as seguintes características:

- *Cluster 1*: indivíduos "próximos" da condição de hipertensos, obesos, idade intermediária e alto consumo de cigarros;
- *Cluster 2:* indivíduos não-hipertensos e de pouca idade;
- *Cluster 3:* indivíduos hipertensos, com idade avançada.

Através da investigação – em nível descritivo, após a *clusterização* – de outras características dos indivíduos, foi observado que a escolaridade dos indivíduos pertencentes ao *cluster 3* é inferior à observada nos demais *clusters.* O *cluster 1* é predominantemente formado por indivíduos do sexo masculino e os *clusters 2 e 3* são formados, em sua maioria, por indivíduos do sexo feminino. Tal resultado foi induzido pelo comportamento das variáveis peso e altura, pois os homens, em média, têm altura superior a das mulheres. Constatou-se também que a renda média dos indivíduos do *cluster 1* é superior a dos demais *clusters.*

Ao se classificar os indivíduos pertencentes a uma população por características que compartilham e, portanto, os identificam como pertencentes a *clusters* desta população, pode-se utilizar esta caracterização dos *clusters* para buscar associações causais de doenças ou ainda idealizar intervenções de saúde que possam ser mais facilmente aplicáveis a cada *cluster.* Assim, por exemplo, em relação aos resultados do estudo presente, poder-se-ia inferir que a obesidade precede a elevação da pressão arterial, ou, ainda, que ações de saúde de caráter preventivo devem ser dirigidas, preferencialmente, aos indivíduos pertencentes ao *cluster* 1, ações de promoção de saúde aos indivíduos do *cluster* 2, enquanto ações assistenciais devem ser dirigidas de forma especial aos indivíduos do c*luster* 3. Devido às diferentes características de escolaridade e de renda dos *clusters*, as intervenções de saúde devem ser adequadas às características sócio-econômicas identificadas.

Referências Bibliográficas

Bezdek, J.C.; Ehrlich, R.; Full, W. (1984). FCM: the fuzzy c-means clustering algorithm, *Computers & Geosciences* 10 (2/3), 91-203.

Bloch, K.V. et al. (1994). Hipertensão arterial e obesidade na Ilha do Governador – Rio de Janeiro, *Arq. Bras. Cardiol.* 62 (7); 17-22.

Cover, T.M; Joy A.T. (1991). *Elements of Information Theory,* Wiley.

Hart, J.T.; Savage, W. (2000). *Tudo Sobre Hipertensão Arterial,* Andrei, São Paulo.

Johnson, R.A.; Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, Prentice Hall.

Klein, C.H.; Souza e Silva, N.A.; Nogueira, A.R.; Bloch, K.V.; Salis Campos, L.H. (1995a). Hipertensão Arterial na Ilha do Governador, RJ, Brasil - I-Metodologia, *Cad. Saúde Pública (Reports in Public Health)*; 11: 187-201.

Klein, C.H.; Souza e Silva, N.A.; Nogueira, A.R.; Bloch, K.V.; Salis Campos, L.H. (1995b). Hipertensão arterial na Ilha do Governador, RJ, Brasil - II-Prevalência, *Cad. Saúde Pública (Reports in Public Health)*, 11:389-394.

Klir, G.J.; Folger, T.A. (1988). *Fuzzy Sets, Uncertainty and Information*, Prentice-Hall.

Mancilha-Carvalho, J.J.; Souza e Silva, N.A.; Carvalho, J.V.; Lima, J.A.C. (1991). Pressão Arterial em Seis Aldeias Yanomami, *Arq. Bras. Cardiol*, 56: 477-482.

Mancilha-Carvalho, J.J.; Souza e Silva, N.A.; Oliveira J.M.; Arguelles, E.; Silva, J.A.F. (1983). Pressão Arterial e Grupos Sociais - Estudo Epidemiológico, *Arq. Bras. Cardiol.* 40: 115-120.