Fatores de Hipertensão usando Redes Neurais Kohonen – SOM

Resumo — O Mapa Auto-organizável de Kohonen (SOM) é um tipo especial de Rede Neural com aprendizado não-supervisionado. Em um mapa Auto-organizável, os neurônios são colocados em nós de uma grade que é geralmente uni ou bidimensional. No presente estudo utiliza-se essa 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. Obteve-se o banco original aplicando o questionário domiciliar da pesquisa, que apresentou medidas objetivas representando: informações demográficas, características constitucionais e socioeconômicas, hábitos de vida, controle da pressão arterial e consumo de medicamentos. Foram entrevistados 1270 casos usando amostragem domiciliar randômica de conglomerados em dois estágios, estratificada, pela classe socioeconômica. Neste estudo, após a mineração de dados, rotulam-se os neurônios de saída para indicar os clusters que representam. Utiliza-se a técnica de Redes Neurais, Kohonen - Som, cuja aplicação permite a obtenção de grupos de indivíduos que apresentem comportamento semelhante em um conjunto de atributos. Com a informação de cinco características, os resultados indicam a formação de três clusters homogêneos internamente e significativamente diferentes entre si, conforme testes estatísticos. Separar os indivíduos de uma população em cluster propõe, por exemplo, identificar e descrever 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 diferentes políticas públicas de intervenção.

Palavras Chave: análise de cluster, hipertensão, mapas auto-organizáveis, redes neurais.

I. **INTRODUÇÃO**

Efetuar comparações no espaço de atributos e recuperar padrões similares é uma tarefa importante, porém não é um problema simples, pois um SOM é inerentemente não-linear. As redes neurais representam uma tecnologia útil na resolução de problemas não-lineares de alta dimensionalidade, além de serem modelos neurais que se aproximam bastante das características do cérebro humano. Neste contexto, é proposta a utilização de redes neurais que aprendam a reconhecer grupos de padrões similares no espaço de atributos. O sistema proposto, os ˜Mapas Auto-Organizáveis de *Kohonen*, é um sistema baseado no aprendizado competitivo e não-supervisionado [9], [10]. A rede neural não supervisionada, Mapa Auto-organizável de *Kohonen*, fundamenta a formação de agrupamentos, ou *clusters*, de indivíduos, de uma população, que apresentem comportamento semelhante em relação a determinadas características mensuradas. O Mapa Auto-organizável de *Kohonen* (**SOM**) divide o espaço de atributos em diferentes regiões. Cada região pertence a um grupo específico que é caracterizada por vetores de atributos dos indivíduos entrevistados. O mapa de características resultante representa o espaço de atributos através das regiões que os representa.

As Doenças Cardiovasculares (DCV) são a principal causa de morte no País. No Rio de Janeiro em 1998 e ainda hoje, as DVC representaram 30% do total de óbitos, o triplo do percentual de óbitos ocasionados pela segunda causa (câncer). Entre as DVC, 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% no 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 aposentados. Em estudo realizado na Ilha do Governador, no Rio de Janeiro [1] em amostra domiciliar randômica estratificada por classe sócio-econômica encontramos uma prevalência de HA de 38%.

O presente trabalho tem como objetivo extrair informações, do banco de dados obtido a partir 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 utilizaremos as técnicas baseadas em Redes Neurais Auto Organizáveis utilizando algoritmos *Kohonen - SOM* [10] para determinar como os padrões de entrada estão distribuídos em categorias.

Conforme HAYKIN [10] as redes neurais artificiais foram concebidas a partir do conhecimento de que o cérebro processa as informações de maneira inteiramente diferente do processamento encontrado nos computadores digitais convencionais. O principal objetivo almejado com a estrutura de funcionamento de uma rede neural artificial e com algoritmos de aprendizagem é a capacidade de *generalização*. A generalização se refere ao

fato de a rede neural produzir saídas adequadas para entradas que não estavam presentes durante o treinamento (ou aprendizagem). Esta capacidade de processar informação torna possível para as redes neurais resolver problemas complexos, alguns intratáveis por meios convencionais.
A seção 3 descreve o funcionamento das redes neurais deste sistema. São apresentadas, ainda, o modelo de mapeamento de características, a teoria da implementação no treinamento da rede de *Kohonen* e a técnica de representação do espaço de atributos com o "mapa contextual", que mostra a clusterização. Na seção 4 são apresentados alguns resultados preliminares que mostram a eficiência dos vetores de atributos extraídos com o SOM no reconhecimento de classes.

II. **BANCO DE DADOS**

O banco de dados original foi construído no inquérito domiciliar da população da Ilha do Governador (I.G.) [5]. 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 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 seis variáveis selecionadas do banco de dados original.

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
Renda	132,88	289,22	0,00	3653,93	437,05

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

Nesta Tabela 1, nota-se que as características medianas de um indivíduo amostrado são: 41 anos de idade, 65 Kg de peso corporal, 1,62m de altura, pressão arterial diastólica de 80 mm Hg e rende igual a 132,88 dólares.

Em relação a outros dados básicos e à escolaridade dos entrevistados, observa-se na Tabela 2 que 55,6% da amostra é composta por mulheres e mais da metade por indivíduos sequer concluiu o segundo grau. A prevalência de hipertensos (PA $\geq$ 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.

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%

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

Na seleção dos atributos foi usada uma regressão logística, onde a variável dependente foi ser ou não hipertenso, confirma como variáveis mais significativas: idade, sexo, peso, altura, pressão arterial sistólica e diastólica, renda familiar, escolaridade, consumo de cigarros.

Salienta-se, entretanto, que uma possível associação ou relação causal de variável de interesse – hipertensão arterial – com as variáveis contidas no banco de dados não pode ser identificada com uma simples descrição global. Em Bloch [6], 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.

III. ANÁLISE DE CLUSTER

A análise de *cluster* é uma técnica exploratória de dados que tem por objetivo formar agrupamentos de objetos semelhantes a partir de um banco de dados. O conceito de *clusterização* difere do conceito de classificação no sentido que a análise de *cluster* é mais ’primitiva’, na qual nenhuma suposição é feita a respeito dos grupos, assim como seu número e estrutura. Os clusters são obtidos por intermédio da aplicação dos conceitos de similaridade e distância [7].

No contexto deste trabalho, empregam-se vetores no espaço p-dimensional e a medida de distância adotada, entre dois objetos, é a medida euclidiana e de *Manhattan*, caracterizadas nas Equações (1) e (2), que definem a distância entre dois vetores p-dimensionais.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + ... + (x_p - y_p)^2} \qquad (1)$$

$$d\,(x\,,\,y\,) = \sum |X - Y| \qquad (2)$$

IV ALGORITMO *SOM* de KOHONEN

As Redes Auto Organizáveis, por exemplo, *Kohonen*, baseadas em aprendizado competitivo, destacam-se como um bom algoritmo [14], em tarefas de clusterização.

Na clusterização *Kohonen* - SOM o indivíduo é apresentado à rede para que agrupamentos de indivíduos similares segundo alguns atributos sejam efetuadas.

Cada um dos n indivíduos – 1231, no caso deste estudo – tem associado a si um vetor $\mathbf{x} = [x_1\ x_2\ x_3\ x_4\ x_5\ ]$ contendo os cinco atributos abaixo:

x_1 – *idade do entrevistado*
x_2– *altura do entrevistado*
x_3– *peso do entrevistado*
x_4 – *pressão sistólica*
x_5 – *pressão diastólica*

A técnica de *Kohonen* - SOM é empregada para que os n vetores $\mathbf{x_j}$ sejam agrupados em **c** *clusters*, cada um com seu centróide.
O algoritmo propriamente dito pode ser descrito através dos passos seguintes:

1. Inicialização dos pesos da rede com valores baixos (0.01 a 0.1) escolhidos aleatoriamente. Ajuste inicial do raio de vizinhança, que poderá começar com a metade do diâmetro da rede e ir diminuindo linearmente.

2. Inserção do padrão de entrada.

3. Cálculo das distâncias de cada saída de acordo com a Equação (3) dada por

$$d_j = \sum_{i=1}^{N} (x_i(t) - w_{ij}(t))^2 \qquad (3)$$

onde d_j é distância entre a saída do neurônio j com a entrada, N é número de entradas, $\mathbf{x_i(t)}$ é o vetor de entrada no tempo t e $\mathbf{w_{ij}(t)}$ é o peso da conexão do neurônio de entrada i para o neurônio j no tempo t.

4. Seleção da menor distância

5. Atualização dos pesos do neurônio com a menor distância (neurônio vencedor) e seus vizinhos, definidos pelo raio de vizinhança. Isto é feito segundo a equação (4)

$$w_i(t+1) = w_i(t) + \alpha(t)\left[x(t) - w_i(t)\right] \qquad (4)$$

onde $\alpha(t)$ e taxa de aprendizado no tempo *t*. A taxa $\alpha(t)$ começa com valor próximo de 1 e decresce monotonicamente conforme a Equação (5).

$$\alpha(t) = 0.9\ (\ 1 - t / 1000) \qquad (5)$$

Os neurônios que não pertencem à vizinhança do vencedor não devem ter seus pesos atualizados.

6. Repetir a partir do passo 2 ate a formação do mapa de característica estar completo, isto e, ate que não sejam observadas modificações significativas no mapa de características.

Quando o algoritmo é inicializado através da escolha de conjunto de valores (geralmente aleatórios) para os pesos iniciais da rede SOM, em seguida inicia-se o procedimento iterativo, estimando-se os vetores centróides e atualizando-se o vetor de pesos $\mathbf{w_{ij}(t)}$, onde IW representa uma matriz SxR, onde S é o número de neurônios e R a dimensão dos dados de entrada.

Para avaliar o desempenho da rede construída a partir da função *Newsom* do *Matlab*, simulamos várias redes alterando os parâmetros? tamanho da rede, topologia, função de distância e o número de *epochs* de treinamento, considerando fixos os outros.

V DISCUSSÃO E RESULTADOS

Na seleção de dados foram excluídos os registros, 3,07% do total, que tinham alguma das variáveis significativas não preenchidas ou com valores inválidos. Foram processados 1231 registros dos 1270 registros iniciais.

Dado que a variável consumo de cigarros apresenta vários valores iguais a zero, foi necessário efetuar de início a normalização dos atributos, de forma a se ter uma variância igual a 1 [14].

Foram realizados experimentos usando o arquivo original e o mesmo normalizado, usando-se todos os atributos, variando o tamanho da rede, a topologia da vizinhança, a função de distância e o número de *epochs* de treinamento.

Os melhores resultados obtivemos quando suprimimos os atributos sexo, escolaridade, renda e consumo de cigarros.

As redes com menos de 500 vezes o número de processadores não apresentaram bons resultados, não permitindo uma discriminação dos indivíduos nos clusters independentemente do número de clusters.

Com diversas tentativas, procurou-se encontrar o número de neurônios necessários que permitissem definir os clusters adequadamente de modo que a rede convergisse.
Observa-se uma evolução da especialização da rede em clusterizar os diferentes padrões de treinamento, ao criar a rede mantendo as variáveis idade, peso, altura, pressão sistólica e pressão diastólica, tendo sido normalizados os padrões de entrada. O resultado obtido foi uma rede de topologia quadrada (*gridtop*), de tamanho 20X20, distância de *Manhattan* e 200.000 *epochs* para as 5 variáveis consideradas..

Através da análise das cinco características orgânicas dos indivíduos, obtivemos 3 grupos cujas principais estatísticas estão dispostas na Tabela III.

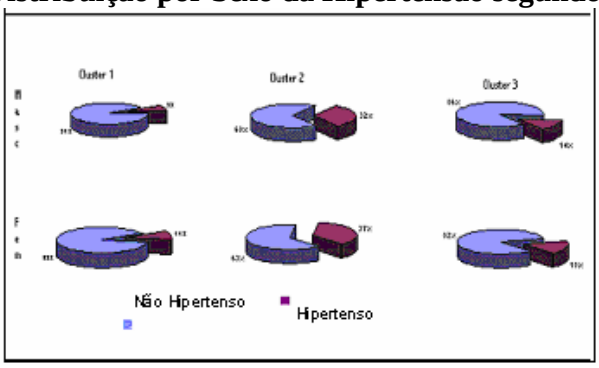
Os clusters foram rotulados com os números 1,2 e 3. Os números entre parênteses indicam o número de indivíduos em cada cluster.

Tabela III
índice de massa corpórea **Estatísticas descritivas das principais variáveis nos 3 Clusters**

	Idade	Cigarros	Peso	Altura	Sist2	Dias2
Cluster 1 (395)	40,57	16,28	77,56	170,81	135,76	84,17
Cluster 2 (4759)	33,81	7,03	57,16	156,17	111,65	79,47
Cluster 3 (361)	58,21	7,11	62,01	151,92	151,52	88,24
Média Geral	43,54	9,51	66,41	162,59	131,12	81,12

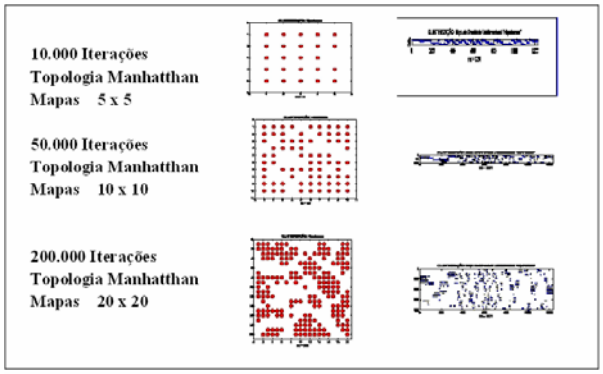
Pode-se observar que os grupos resultantes apresentam características bem definidas. O terceiro cluster agrega os mais idosos e com maior pressão sistólica e diastólica, sendo o segundo colocado em consumo de cigarros e índice de massa corpórea. No segundo agrupamento encontram-se os mais jovens, que consomem menos cigarros e com menor índice de massa corpórea e, portanto, associados a menores medidas de pressão. O primeiro cluster contem os indivíduos mais jovens que os do cluster 3 , mas que fumam mais e que tem a segunda maior medida de pressão. São também indivíduos de maior índice de massa corpórea.

Figura 3 – Distribuição por Sexo da Hipertensão segundo os Clusters



Para avaliar a discriminação do algoritmo SOM, são usados o mapa contextual (clusterização) e o mapa de densidade unidimensional. A figura 4(d) ilustra os resultados desses mapas para três diferentes configurações.

Figura 4(d) – Resultados de redes



6 – CONCLUSÃO

Ao nível de abstração, os testes de grupamento de vetores através de medidas de similaridade, usando Kohonen - SOM obteve uma excelente representação dos padrões usando uma rede 20X20, distância de Manhattan. 200.000 iterações para 5 variáveis do arquivo de entrada normalizado, onde se visualizam 3 clusters. Comparando-se esses resultados com o trabalho de Vellasco, M. M. et al, [12], assinala-se que esses são consonantes com os mesmos. A partir desse resultado podemos concluir que SOM pode considerada uma excelente técnica de clusterização recomendando seu uso para mineração de dados. Entretanto, é necessário um estudo mais detalhado da eficiência do clusterizador, dada a grande possibilidade de variação dos parâmetros das redes. Comparações com outros clusterizadores, exceto Fuzzy Clustering, não foram feitas porque a preocupação maior deste trabalho foi com a implementação da extração de atributos e do classificador neural. Comparações com outras abordagens, bem como o desenvolvimento de um sistema completo de busca e recuperação de informações visuais por padrões será assunto de trabalhos futuros.

VII. REFERÊNCIAS BIBLIOGRÁFICAS

[1] C.H. Klein, N.A. Souza e Silva, A.R. Nogueira, K.V. Bloch, L.H. Salis Campos, Hipertensão Arterial na Ilha do Governador, Rio de Janeiro – Brasil – II – Prevalência, *Cad. Saúde Pública (Reports in Public Health),* 11:389-394, 1995.

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

[3] J.J. Mancilha-Carvalho, N.A. Souza e Silva, J.M. Oliveira, E. Arguelles, J.A.F Silva, Pressão Arterial e Grupos Sociais – Estudo Epidemiológico, *Arq. Brás. Cardiol.* 40:115-120, 1983.

[4] J.T. Hart, W. Savage, *Tudo Sobre Hipertensão Arterial*, Andrei, São Paulo, 2000.

[5] C.H. Klein, N.A. Souza e Silva, A.R. Nogueira, K.V. Bloch, L.H. Salis Campos, hipertensão Arterial na Ilha do Governador, Rio de Janeiro – Brasil – I-Metodologia, *Cad. Saúde Pública (Reports in Public Health)*; 11: 187-201, 1995.

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

[7] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 4ª ed, Prentice Hall, 1998.

[8] T.M. Cover, A.T. Joy, *Elements of Information Theory*, John Wiley and Sons, New York, 1991.

[9] Braga, A. P., Ludemir T. B. e Carvalho A. C. P. L. F. (1999):, Redes Neurais Artificiais –Teorias e aplicações.

[10] Haykin, S. Redes Neurais, Princípios e Prática. 2 ed., Bookman, 900 pág., 2001.

[11] Kohonen, T. The Self-Organizing Map, Proceedings of the IEEE, vol. 78, no. 9, p. 1464 – 1480, 1990.

[12] Vellasco, M. M. et al., Aplicação de Fuzzy Clustering a Banco de Dados de Amostra Domiciliar da População da Ilha do Governador.

[13]Notas de aula, www.ica.ele.puc-rio.br Acesso: 2005

[14]L. FU, Neural Network in Computer Intelligence, MacGrawHill, 1994.