Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais
pp. 373-377, July 20-22, 1999 - ITA, São José dos Campos - SP - Brazil

# Segmentação Híbrida com Aplicação em Análise de Endividamento Privado Internacional

A.M.D. Monteiro [1, 3], D.D. Carneiro [2], C.E. Pedreira [3]
[1] Núcleo de Pesquisas, Banco Icatu, Rio de Janeiro, Brazil;
[2] Dept. de Economia, PUC-RIO;
[3] Dept de Eng. Elétrica, PUC-RIO, C.P. 38063, CEP 22452-970, Rio de Janeiro, Brazil;
andremonteiro@icatu.com.br, dionisio@econ.puc-rio.br, pedreira@ele.puc-rio.br

**Abstract**

The main goal of this paper is to propose procedures for segmentation by combining statistical and connectionist schemes. The procedures are tested in two cases. The first one is a bidimensional synthetic example. The second is real economic problem: the relationship between private debts and some macroeconomic variables of thirty-nine countries are discussed in some detail. We concluded that clustering performance is improved by taking advantages of specific properties and capacities of each method.

## 1. Introdução

A principal finalidade de procedimentos de segmentação é obter grupos bem definidos e compactos. Métodos estatísticos clássicos [1] têm obtido razoável sucesso em algumas aplicações, embora apresentem bastante sensibilidade à escolha das condições de inicialização. O Mapa Auto Organizável proposto por Kohonen [2][3][4] produz um ajuste seletivo dos neurônios criando um mapa topográfico dos padrões de entrada. Ainda que esta técnica não tenha sido originalmente concebida para classificação de padrões ou segmentação, é possível tirar-se proveito das suas propriedades de auto organização para tais aplicações.

Neste artigo, serão propostos procedimentos para segmentação que combinam o Mapa Auto Organizável de Kohonen com técnicas estatísticas. Duas aplicações são exploradas: um exemplo artificialmente criado e um problema real. A primeira mostra as potencialidades dos procedimentos propostos através de um exemplo bidimensional controlado. Na segunda serão estudadas as experiências de trinta e nove países com as relações entre endividamento privado e algumas variáveis usualmente utilizadas para avaliação de desempenho macroeconômico. Na segunda seção são descritos procedimentos usados para segmentação. Além do algoritmo de Kohonen, apresenta-se um método híbrido no qual é explorada a redução de dispersão dos segmentos. Na seção 3 são apresentados experimentos controlados com o objetivo de explorar a potencialidade de algoritmos que associam a auto organização de Kohonen a minimização de dispersão de agrupamentos. Na quarta seção estes procedimentos são aplicados em um problema de análise de endividamento internacional. A última seção ficou reservada para comentários finais.

## 2. Procedimentos Para Segmentação

Uma estratégia usual para resolver problemas de segmentação consiste em minimizar uma função de custo associada às dispersões dos segmentos. Em geral, é interessante associar, a cada grupo, um protótipo com localização o mais central possível no grupo. Vale ressaltar que o mapa auto-organizável de Kohonen gera, na convergência, um conjunto de protótipos que não necessariamente satisfaz a esta propriedade [5]. Mais do que isso, a minimização da dispersão dos segmentos não é um dos objetivos deste algoritmo. Por outro lado, sua capacidade de aproximar, preliminarmente, a função densidade de probabilidade dos padrões de entrada pode ser de grande utilidade para problemas de segmentação.

A motivação do procedimento proposto a seguir é a de tirar partido da capacidade de aproximação preliminar do Mapa Auto Organizável e subseqüentemente minimizar a dispersão média intra-segmentos. Desta forma o procedimento dá-se em duas etapas: na primeira, aplica-se o algoritmo de Kohonen; na segunda, minimiza-se a dispersão intra-segmentos. Vale notar que na primeira etapa, não se busca convergência fina do Mapa de Kohonen. Assim, para cada método de minimização da dispersão empregado, tem-se um procedimento diferente. Neste artigo, dois métodos são utilizados: o algoritmo de K-means e uma heurística proposta na seqüência.

### 2.1. O Mapa Auto Organizável de Kohonen

O objetivo básico do Mapa Auto Organizável proposto por Kohonen é agrupar n elementos de um conjunto de padrões de entrada, X, em J neurônios (ou dependendo do contexto, em grupos, ou em segmentos) distribuídos em uma malha de 1, 2 ou 3 dimensões. A "comunicação" entre os ambientes de entrada (subespaço natural dos padrões de entrada) e o de saída do algoritmo (malha) é feita por protótipos. O protótipo

(também chamado de peso no contexto de redes neurais) é um elemento iterativamente construído pelo algoritmo e inserido no conjunto X. A cada neurônio associa-se um vetor que será interpretado como protótipo.

O algoritmo proposto por Kohonen pode ser dividido em quatro etapas básicas: (i) Cálculo das distâncias aos J protótipos de um elemento sorteado $X_k$ do conjunto de padrões de entrada; (ii) Comparação dos valores das J distâncias e reconhecimento do menor, ou seja, aponta-se o protótipo mais próximo de $X_k$ no sub-espaço de saída; (iii) Ativação, por uma rede interativa , simultaneamente, do neurônio vencedor e da sua vizinhança; (iv) Diminuição gradativa, através de processo adaptativo, da distância do neurônio vencedor e da respectiva vizinhança a $X_k$.

A necessidade de mensurar distância impõe a escolha de duas métricas: uma para o ambiente de entrada (d) e outra para a malha (d*). Define-se vizinhança (V) de raio $R_v$ do neurônio j como o conjunto de neurônios cuja distância a j é inferior ou igual a $R_v$ na malha:

$$V = V(R_v) = \{N: d^*(C_j, C_i) \leq R_v\}$$

O conjunto de neurônios $V(R_v) = r$ é dito r-ésima vizinhança. Embora métricas d* diferentes possam gerar vizinhanças diferentes, resultados empíricos mostram que o resultado final do Mapa não é afetado de forma significativa pela escolha da métrica da malha.

As tarefas de acionar o neurônio vencedor e sua vizinhança e de fixar as parcelas para a diminuição da distância destes ao elemento apresentado cabem à função de vizinhança centrada no neurônio vencedor, $\Im_{C_v}$ : o protótipo vencedor é atualizado pelo fator unitário e os demais protótipos também podem ser atualizados por fatores diferentes de zero, com valores proporcionais às distâncias, medidas por d*, dos respectivos neurônios ao neurônio vencedor.

Estabelecida a forma da função de vizinhança, define-se o raio de atualização ($R_a$) como o parâmetro que determina até qual vizinhança, em relação ao neurônio vencedor, será efetuada a atualização dos protótipos. O raio $R_a$ pode ser variante no tempo. Deste modo, função de vizinhança centrada no protótipo vencedor v em t, contador do número de iterações, é dada pela seguinte relação:

$$\Im_{C_v} = \Im_{C_v}(X(t),\ R_a(t))$$

Os protótipos são agrupados na matriz U, onde na coluna j está o protótipo associado ao neurônio j. O treinamento então, dá-se em quatro passos:

P1. Sortear aleatoriamente uma entrada $X_k$, tal que $X(t)=X_k$;

P2. Encontrar o neurônio vencedor v tal que:
$v = \arg_j \min d(X(t), U_j(t))$ , j = 1,2, ... J;

P3. Atualizar a matriz de protótipos U através de:
$U(t+1) = U(t) + \gamma(t).\Im_{C_v}(X_k, R_a(t)).[X_k - U_v(t)]$
onde $\gamma(t)$ é taxa de aprendizado em t,

P4. Interromper o processo quando não forem detectadas alterações significativas nos protótipos.

Seguindo-se a abordagem variacional de Likhovidov [5], demonstra-se que, caso o protótipo convirja, ele não o faz, necessariamente, para a posição central do grupo.

O Mapa Auto Organizável apresenta duas propriedades importantes. A primeira é que ele aproxima a função densidade de probabilidade dos padrões de entrada, p(x). A transformação que ele realiza é capaz de capturar variações nas estatísticas da distribuição dos elementos do conjunto de entrada. Devido ao sorteio aleatório - passo 1 do treinamento - maior número de neurônios é deslocado para cobrir regiões de alta densidade de probabilidade. Em outras palavras, regiões do domínio dos dados de entrada com maiores probabilidades associadas ocupam mais neurônios na malha. Portanto, a resolução para estas regiões é maior: a malha tem maior habilidade em diferenciar elementos de entrada que estão em zonas de maior probabilidade p(X). A segunda propriedade é a manutenção da topologia do sub-espaço de saída. Em [6] é apresentada uma prova de ordenação de variável unidimensional em malha de uma dimensão.

### 2.2 Métodos para redução de dispersão

A tarefa de minimização de dispersão intra-segmentos ficará a cargo de dois métodos. O primeiro método é o K-means. O segundo é uma heurística proposta a seguir. A dispersão a ser minimizada é definida como a média do desvio padrão de cada segmento ponderada pelos respectivo número de elementos. O que se busca é atribuir um peso maior para os grupos mais densos.

A heurística proposta é composta por 'loopings' Locais e Globais. Os loopings Locais objetivam achar o melhor estimador de posição para o grupo em questão, de acordo com um critério exposto em seguida. Nestes loopings, são gerados candidatos a protótipo para o grupo, enquanto que para todos os demais grupos nada muda. O looping Global é composto de uma rodada de loopings locais para cada um dos grupos. No looping Global os candidatos a protótipos identificados como sendo os melhores para cada um dos grupos são implementados, consolidando o processo. Os primeiros candidatos a protótipos são aqueles gerados pelo algoritmo de Kohonen. A seguir, apresentam-se os passos da heurística:

P1: Gerar os segmentos, utilizando-se os atuais candidatos a protótipos, agrupando os dados de acordo a menor distância aos protótipos;

P2: Escolher um dos grupos gerados em P1;

P3: Calcular o estimador de posição deste grupo (e.g. média, mediana), e nomea-lo como o novo candidato a protótipo;

P4: Re-alocar todos os dados usando o critério de proximidade dos protótipos;

P5: Calcular o desvio padrão para o grupo escolhido. Conservar este resultado para o teste de estabilidade do passo 7.

P6: Calcular a dispersão da segmentação. Reter este resultado para comparação no passo 8;

P7: Se a estabilidade local não foi atingida, isto é, se o passo 5 não produziu resultado idêntico em duas iterações consecutivas, retorne ao passo 3.

P8: Escolha o candidato a protótipo que corresponde a menor dispersão da segmentação calculada no passo 6.

P9: Retornar ao passo 2 sem trocar nenhum dos candidatos a protótipo. Recomeçar este passo até que todos os grupos sejam visitados.

P10: Escolher os novos candidatos a protótipo para todos os grupos usando o critério de minimização da dispersão da segmentação (passo 6) até que a Estabilidade Global seja alcançada. Por Estabilidade Global compreende-se que nenhum dos candidatos a protótipo (passo7) é trocado em duas rodadas Globais consecutivas.

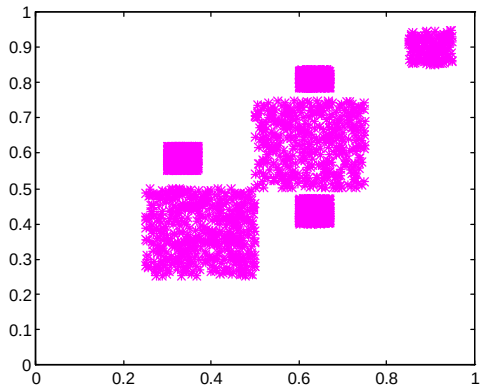
Nota-se que a fase de loopings Locais pode ser processada de modo paralelo. A heurística proposta guarda semelhança com o método de K-means. São duas as diferenças entre eles: na heurística, a troca de protótipos é condicionada à redução da dispersão total (passo 8) e os candidatos a protótipos não são apenas as médias, podem ser qualquer estimador de posição.

### 3. Um Experimento Controlado

Nesta seção serão apresentados os resultados numéricos de um experimento controlado com o propósito de ilustrar o comportamento dos procedimentos propostos.

Com um total de 4075 pontos, foram gerados 6 grupos sintéticos (i.e. dados artificiais). Todos obedecem a uma distribuição uniforme dentro de polígonos -- tamanhos e densidades diferem. Estes segmentos encontram-se representados graficamente na figura 1.

**Figura 1**: Seis segmentos gerados artificialmente




Este problema bidimensional foi abordado de quatro formas diferentes: Kohonen + heurística; Kohonen + K-means; Kohonen puro; e K-means puro. A distância Euclidiana foi escolhida como métrica. As segmentações resultantes estão nas figuras 2 a 4, onde as linhas contínuas representam as fronteiras geradas entre os grupos.

**Figura 2**: Segmentação gerada por Kohonen + heurística

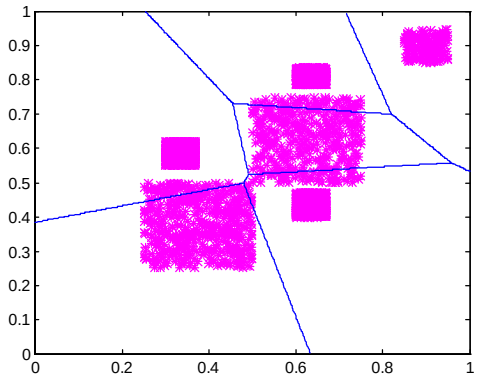



**Figura 3**: Segmentação gerada por Kohonen+K-means

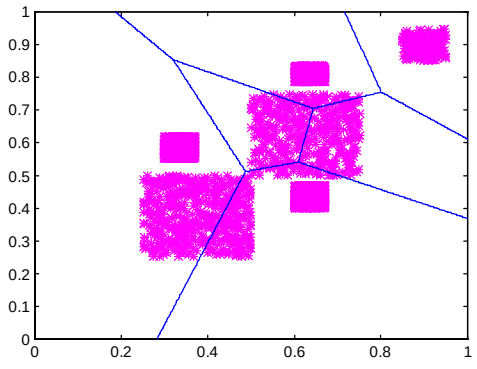
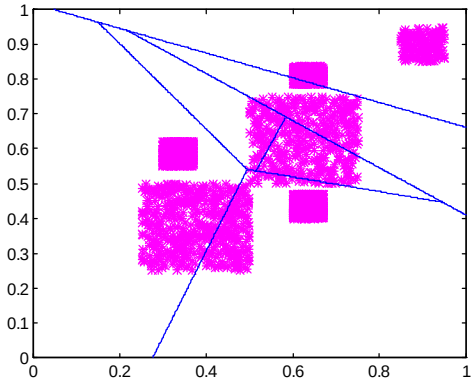


**Figura 4**: Segmentação gerada por Kohonen

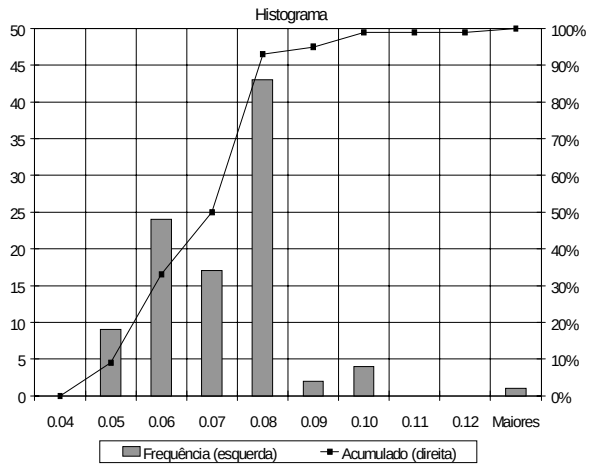



O procedimento Kohonen + heurística foi capaz de reconhecer todos os 6 segmentos, com dispersão média de 0.0476. O procedimento Kohonen + K-means não foi capaz de distinguir dois dos grupos e dividiu em duas partes um dos segmentos; obteve dispersão de 0.0702. A aplicação apenas de Kohonen gerou resultados bastante pobres, com dispersão de 0.0894. Com relação a dispersão obtida, pode-se afirmar que o procedimento Kohonen + K-means alcançou um ganho de 21% se comparado com o método de Kohonen puro; enquanto Kohonen + heurística, 47%.

A comparação entre as técnicas acima e K-means deve ser feita com base nas probabilidades *ex-pos t*das diversas segmentações geradas pelo último, já que seu desempenho é conhecido por ser fortemente afetado pelas condições de inicialização. O método de K-means foi rodado independentemente 100 vezes, com inicialização randômica. A figura 5 ilustra os resultados.

**Figura 5**: Resultados de K-means

Fazendo a comparação entre K-means e Kohonen + heurística, as segmentações produzidas foram idênticas em 9% das vezes. Por outro lado, o procedimento gerou segmentação superior a 91% daquelas produzidas por K-means.




Ambos os procedimentos (Kohonen + heurística e Kohonen + K-means) apresentaram resultados superiores a Kohonen ou K-means quando aplicados individualmente. Esta fato parece indicar que a combinação da organização topologia para inicialização da minimização da dispersão produz um ganho real nos processos de segmentação.

## 4. Análise de Endividamento Privado Internacional

Nos anos 90, observou-se uma explosão de fluxos internacionais de capitais na direção de mercados emergentes. Além disso, uma larga variedade de desempenho macroeconômico de diferentes países passou a desafiar as classificações usuais, baseadas na simples avaliação de déficits fiscais. Além das dificuldades associadas com a comparação de diferentes medidas de variáveis fiscais, causalidade entre déficit fiscal e desempenho macroeconômico global é incerta. A dispersão do endividamento privado e a sua aparente falta de associação direta com as medidas usuais de desempenho macroeconômico (e.g. crescimento, renda per capita, déficit externo e inflação) sugere que esta pode ser útil como uma variável adicional na classificação do desempenho macroeconômico dos países a médio prazo.

O objetivo desta seção é analisar a relação entre endividamentos privados como percentagem do PIB e o desempenho macroeconômico. Este último é observado pelas seguintes variáveis (a fonte principal para estas foi o International Financial Statistics, IFS, -- March 1997 do Fundo Monetário Internacional): taxa inflacionária, superávit em conta corrente como porcentagem do PIB, renda per capita e taxa de crescimento econômico. As variáveis foram manipuladas e normalizadas como médias anuais no período entre 1991 e 1995.

São duas as motivações para aplicar os procedimentos descritos anteriormente neste problema. Primeiro, a estrutura é multivariada e, portanto, a sua visualização é não trivial. Segundo, os detalhes e idiossincrasias dos países danificariam a análise global se cada um destes tiver as sua diversidade escrutinadas isoladamente. Deste modo, segmentação parece ser um procedimento apropriado. Nota-se que não há intenção em criar uma função que gere valores para endividamentos privados como uma porcentagem do PIB a partir de quatro variáveis, nem tão pouco em decompor a variância total em componentes.

Foram selecionados trinta e nove países de um total de 160 disponível no IFS. O critério utilizado para esta seleção baseou-se na relevância destes no cenário mundial e na qualidade dos dados. Os países escolhidos foram: Argentina, Austrália, Áustria, Bélgica, Bolívia, Canadá, Chile, China, Colômbia, Dinamarca, Egito, Finlândia, França, Alemanha, Grécia, Holanda,

Hungria Índia, Indonésia, Israel, Itália, Japão, Malásia, Marrocos, México, Noruega, Paraguai, Peru, Portugal, Singapura, Coréia do Sul, Espanha, Suíça, Tailândia, Turquia, Reino Unido, Uruguai, EUA e Venezuela.

A distribuição de países por continente é: Ásia, 9; Europa, 16; América Norte, 2; América Latina, 9; África, 2; e Oriente Médio, 1. Dezoito são países desenvolvidos. Vale ressaltar que esta seção é parte de um projeto de pesquisa em dívidas privadas que cobre dados desde 1981. No início dos anos 80, os dados dos países do Leste Europeu não estavam disponíveis no IFS, e por esta razão não aparecem aqui. O Brasil não foi incluído por causa da sua enorme taxa inflacionaria verificada neste período.

Após análise preliminar, concluiu-se que quatro grupos representariam adequadamente a estrutura dos padrões de entrada. A organização em grupos (C1 a C4) corresponde a distâncias unidimensionais de modo que C1 está, por exemplo, mais próximo a C2 que a C4. O problema foi abordado por Kohonen + K-means, Kohonen + heurística, e Kohonen e K-means isoladamente. A segmentação mais coerentes do ponto de vista macroeconômico e, também, que produziu menor dispersão foi aquela gerado pelo procedimento Kohonen + heurística. Os resultados são descritos a seguir.

C1: Peru, Turquia, Uruguai, e Venezuela;
C2: Argentina, Bolívia, Chile, China, Colômbia, Egito, Grécia, Hungria, Índia, Indonésia, Israel, Coréia, Malásia, Marrocos, México, Paraguai, e Tailândia;
C3: Alemanha, Austrália, Áustria, Canada, Dinamarca, Finlândia, França, Itália, Portugal, Espanha, EUA e UK;
C4: Bélgica, Holanda, Japão, Noruega, Singapura e Suíça.

**Tabela 1**: Médias das variáveis por grupo de países

| Grupos | Credito/ PIB | Inflação | Renda Per Capita | Cres-cimento | CC/ PIB |
|---|---|---|---|---|---|
| C1 | -1.08 | 2.56 | -0.87 | 0.01 | -0.17 |
| C2 | -0.33 | 0.00 | -0.81 | 0.48 | -0.51 |
| C3 | 0.35 | -0.54 | 0.74 | -0.56 | -0.12 |
| C4 | 0.87 | -0.61 | 1.29 | -0.30 | 1.52 |

Todas as variáveis discriminam C1+C2 de C3+C4, exceto a conta-corrente como percentagem do PIB. Os primeiros dois grupos apresentam os menores valores para a variável crédito e renda per capital e os maiores para inflação. A linha divisória mais forte é a renda per capita: apenas um único país em C1+C2 apresenta esta variável acima da média amostral (Israel). A caracterização de C1 pela alta inflação é notável: ele é composto exatamente pelos países que experimentaram as mais altas taxas no período. A variável que melhor discrimina C3 de C4 é a conta-corrente como percentagem do PIB: todos os seis países têm os maiores valores da amostra (desconsiderando o Egito).

A consistência econômica e estatística da segmentação acima indica que, realmente, os países analisados passaram por experiências similares. Existem, conforme suspeitado, associações entre o crédito ao setor privado com percentagem do PIB e algumas das variáveis usuais de desempenho macroeconômico. Sua associação é muito forte com inflação (negativa) e renda per capita (positiva). Agrupando C1 com C2 e C3 com C4, encontra-se relação, também, com o crescimento econômico (negativa). Com a variável conta-corrente como percentagem do PIB, não foi identificada associação.

Vale notar que os grupos trazem uma clara componente regional, apesar desta informação não ter sido dada aos métodos de segmentação.

## 5. Conclusão

Neste artigo foram apresentados procedimentos para segmentação que combinam o Mapa Auto Organizável de Kohonen com ferramentas de redução de dispersão total. Os procedimentos mostraram-se mais eficientes que as técnicas aplicadas isoladamente tanto no exemplo artificialmente construído quanto no problema real quando foram analisadas as experiências de vários países com o crédito ao setor privado e as demais variáveis usuais de desempenho macroeconômico.

## Referências


[1] W.R. Dillon and M. Goldstein. Multivariate Analysis, John Wiley & Sons, 1984.
[2] T. Kohonen. Self-organized formation of topologically correct feature maps, Biological Cybernetics 43, 59-69, 1982.
[3] T. Kohonen. Self-organizing Maps, Springer Verlag, 1995.
[4] S.Kaski e T. Kohonen . Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world, In: Neural Networks in the Capital Markets, World Scientific, 1996;
[5] V. Likhovidov. Variational Approach to Unsupervised Learning Algorithms of Neural Networks, Neural Networks, vol 10, Nº 2, 273-289, 1997.