



Revista Pensamento Contemporâneo em
Administração

E-ISSN: 1982-2596

jmoraes@id.uff.br

Universidade Federal Fluminense
Brasil

Guimarães De Pinho, Anderson
MINERAÇÃO DE DADOS COM MAPAS DE KOHONEN: UMA ABORDAGEM NO
SETOR FINANCEIRO
Revista Pensamento Contemporâneo em Administração, vol. 2, núm. 1, mayo-agosto,
2008, pp. 39-49
Universidade Federal Fluminense
Rio de Janeiro, Brasil

Disponível em: <http://www.redalyc.org/articulo.oa?id=441742832004>

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

redalyc.org

Sistema de Informação Científica
Rede de Revistas Científicas da América Latina, Caribe, Espanha e Portugal
Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto



MINERAÇÃO DE DADOS COM MAPAS DE KOHONEN: UMA ABORDAGEM NO SETOR FINANCEIRO

DATA MINING WITH KOHONEN MAPS: AN APPROACH IN FINANCIAL SECTOR

Anderson Guimarães De Pinho

RESUMO

O presente artigo objetivará a aplicação de Mapas de Kohonen na identificação de perfis de clientes, ou clusters, como estratégias de diferenciação no marketing empresarial. Clientes que decidem por investir em ações na Bolsa de Valores de São Paulo (Bovespa) podem optar por ações de empresas nos mais diversos ramos de atividade. Se gestores do marketing souberem explicitar o conhecimento necessário para diferenciar clientes de acordo com o setor em que operam na bolsa de valores, estes poderão incorporar este conhecimento nos processos da empresa, maximizando lucros e resultados. Mapas de Kohonen representam uma classe de modelos de redes neurais denominados mapas auto-organizáveis e seu potencial de aplicação em processos de descoberta de conhecimento representa alguma vantagem competitiva. Em nosso caso de aplicação isto significa separar clientes de acordo com suas concentrações de operações na Bovespa nos 10 possíveis mercados de ações. Os resultados mostrarão que clientes podem ser agrupados em 1 de 6 clusters diferentes, cada um com características bastante singulares.

Palavras-chave: Clusterização de clientes. Mapas de kohonen. Data mining. Redes som. Marketing de relacionamento.

ABSTRACT

This article aimed at applying the Kohonen Maps to identify profiles of customers or clusters, as a strategy of differentiation in marketing business. Customers who decide to invest Money in the Sao Paolo stock exchange may opt to invest in companies of various branches of activity. If marketing managers know the explicit knowledge that differentiate customers according to the sector in which they operate in the stock exchange, they may incorporate this knowledge in the processes of the company, maximizing profits. Kohonen maps represent a class of models of neural networks known as self-organizing maps and its potential of application in processes of knowledge discovery is a competitive advantage. In our case of application this means implementing separate customers according to their concentrations of investments in 10 markets for possible actions. The results show that customers can be grouped into 1 of 6 different clusters, each with very unique characteristics.

Keywords: Cluster of customers. Kohonen maps. Data mining. Som networks. Relationship management.

INTRODUÇÃO

A forma com a quais grandes empresas interagem com seus consumidores mudou drasticamente nos últimos anos em função dos avanços da tecnologia. A expectativa dos consumidores aumentou bastante, e tem se tornado mais difícil satisfazê-los. A cada dia, mais produtos e serviços similares são lançados ao mercado, diversificando ainda mais as opções de escolha dos consumidores. Com isso, torna-se difícil obter a lealdade dos consumidores por uma firma em particular. Da mesma forma, as empresas necessitam reter seus clientes mais rentáveis para sobreviver em mercados competitivos e dinâmicos.

Como resultado, as grandes empresas necessitam entender melhor o comportamento dos seus consumidores e incorporar este conhecimento em seus processos. Mais conhecimento se traduz em uma melhor resposta às expectativas dos consumidores, de forma rápida e eficaz. Resposta esta, quando possível, pró-ativa, antecipando-se aos desejos dos consumidores e as estratégias da concorrência. E por que não inovando?

Para que isto seja realizado, empresas estão coletando e armazenando informações das mais diversas dos seus clientes, como: geografia, demografia, atitudes, psicografia, comportamentais, e, acredito como sendo a mais importante de todas, informação *transacional*. Muitos avanços têm ocorrido na forma com a qual grandes empresas obtêm, organizam e gerenciam estas informações. Cada vez mais sistemas de informação são desenvolvidos para que os processos de uma empresa sejam gerenciados ao nível de cliente, de maneira quase que pessoal e customizada, extraíndo e armazenando sempre que possível informação útil desta interação.

Neste ambiente, clusterizar clientes em grupos o mais homogêneos têm demonstrado uma estratégia eficaz de marketing. Ações de *cross-selling* ou *up-selling* (Berry, p.115) podem ser alinhadas para maximizar a aceitação, ou satisfação, por determinados grupos de indivíduos. Uma vez que identificados estes clusters onde clientes apresentam atitudes/comportamentos o mais semelhantes, fica mais fácil a otimização de processos de marketing, sem que seja necessário customizá-los.

A extração do conhecimento, capaz de clusterizar clientes, de grandes bancos de dados é responsabilidade de uma área da ciência denominada *data mining* ou *mineração de dados*. David Hand et al (2001, p.6) define como:

... a análise de dados observados em grandes bancos de dados, para descoberta de relacionamentos inesperados, resumindo-os em uma nova forma de fácil entendimento e grande utilidade para o dono dos dados.

DATA MINING

Principais objetivos

Muitos autores tratam esta ciência como uma componente da área de aprendizado de máquina e computação adaptativa. O objetivo de construir sistemas capazes de adaptarem-se ao seu meio, e extrair informação útil a cada interação tem atraído pesquisadores em muitos campos da ciência como: engenharia, física, estatística, neurociência, etc. David *apud* Thomas (2001, p.2) ressalta que muitos destes estudos têm convergido para um conjunto de questões envolvendo problemas supervisionados, não supervisionados, e aprendizado por reforço (reinforcement learning).

Trata-se de uma nova disciplina envolvendo estatística, aprendizado de máquina, administração de banco de dados, reconhecimento de padrões, inteligência artificial e outras áreas. Todas elas relacionadas a um objetivo principal: analisar grande volume de dados, o que as torna bastante correlatas.

Sobre aplicações na área comercial, Berry (2004, p.7) afirma que é muito difícil encontrarmos um algoritmo de mineração de dados o qual tenha sido criado primeiramente com fins lucrativos como seria numa aplicação de marketing empresarial. A aplicação de técnicas de data mining num contexto comercial depende muito do tipo do problema, da habilidade e preferência do pesquisador, e da disponibilidade de informação útil.

Sendo assim, em vista o objetivo de clusterizar clientes tratado neste artigo (aplicação comercial), introduziremos uma nova definição sobre os objetivos da Data Mining dada por Berry :

... permitir que a empresa melhore o marketing, vendas, e operações de suporte ao cliente através de um melhor entendimento de seus clientes”.

Técnicas de clusterização

Hair *et al* dividem os métodos de clusterização em hierárquicos e não hierárquicos. Métodos hierárquicos envolvem a hierarquia de estrutura tipo árvore, sendo divididos em aglomerativos ou divisivos. Já os não hierárquicos, simplesmente tentam encontrar o melhor agrupamento de indivíduos, uma vez que o número de cluster tenha sido especificado, através de determinação e adaptação de centróides. Procedimentos de clusterização desta última natureza são freqüentemente chamados de K-means.

Em processos de Knowledge Discover Database (KDD), uma das técnicas com bastante destaque são os Mapas Auto-Organizáveis, ou Self Organizing Map (SOM). Apesar deste último mencionado, tratar-se de um método não hierárquico de clusterização, veremos que em muito difere de procedimentos do tipo K-means.

Berry e Linoff, a exemplo, introduzem uma aplicação em [3] para clusterizar clientes os quais tomam empréstimos bancários. Com isto, a empresa buscava o melhor entendimento destes clientes e possível aumento de participação de mercado pela utilização deste conhecimento. Larose em [6] defende o uso do algoritmo de clusterização BIRCH se comparado com as técnicas K-means e Redes Som, aplicando-o em histórico transacional de clientes os quais fazem compras em shoppings. Já Vesanto e Alhoniemi [7] ressaltam as vantagens das redes SOM em comparação a métodos hierárquicos e não hierárquicos, em especial o algoritmo K-means. Marques em [8], numa aplicação em pesquisa de mercado, conclui que embora as técnicas de redes SOM e K-means apresentem perspectivas bastante diferentes em termos de resultados, podem produzir clusters bastante semelhantes.

Embora muitas técnicas de clusterização dividam pesquisadores na forma de aplicação, acreditamos que as redes SOM apresentam grandes vantagens com relação as demais. São estas vantagens: (1) fornecem uma visualização bidimensional dos clusters a serem analisados; (2) exigem menos esforço computacional; (3) são bastante robustas na presença de dados ruidosos e/ou outliers; (4) não requer que clusters sejam previamente identificados; (5) não estão sujeita a aleatoriedade de comparação de indivíduos, como acontecem em métodos aglomerativos ou separativos. (6) e não fazem suposições na formação dos clusters, a exemplo do K-means, o qual tenta encontrar clusters esféricos. Para mais detalhes

recomendamos a leitura de [7].

Por estes motivos, o presente estudo defenderá as Redes Som ou Mapas Auto- Organizáveis como estratégia de clusterização de clientes. Para tanto, seremos breves em 2.3 quando do detalhamento da técnica. Em 3, descreveremos o problema encontrado para aplicação da técnica. E em 4, daremos continuidade com a apresentação dos resultados obtidos. É importante dizer que dois modelos básicos de mapas auto-organizáveis são tratados na literatura de Redes Neurais, o modelo de Willshaw-von der Malsburg, e o modelo de Kohonen, tendo este segundo recebido mais atenção por não explicar detalhamento neurobiológico (Haykin, p.485). Sobre isso, dizemos que o modelo aqui proposto e estudado na aplicação de clusterização de clientes será o de Kohonen.

Introdução as redes de som

Mapas Auto-Organizáveis trata-se de uma classe especial de redes neurais baseadas na aprendizagem competitiva: neurônios na camada de saída competem entre si para ativação ou disparo, e somente um neurônio é ativado a cada apresentação de padrão. O neurônio ativado é chamado de neurônio vencedor, e somente os pesos associados a este neurônio serão atualizados, “o vencedor ganha tudo”.

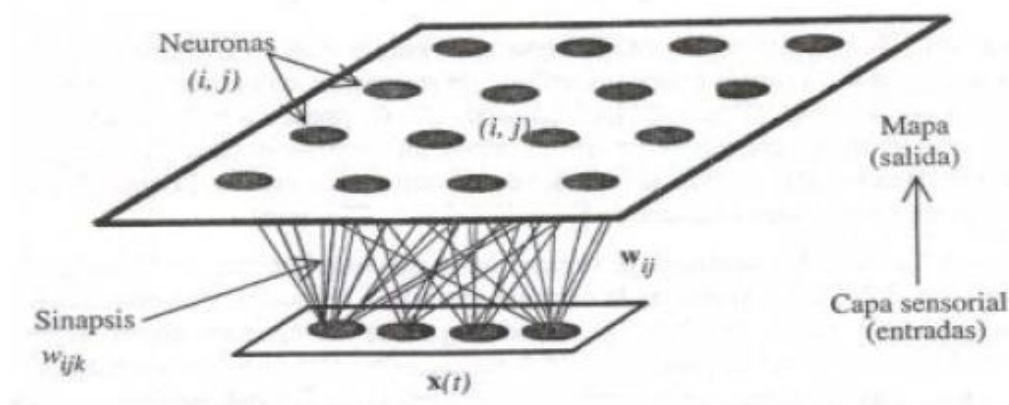
Em uma rede Som, neurônios são dispostos em uma grade uni ou bidimensional. Mapas de dimensionalidade mais alta não são tão comuns, porém possíveis de serem estimados. Os neurônios na camada de saída são ajustados de forma a representar uma classe de padrões semelhantes no conjunto de entrada. As posições dos neurônios numa grade uni ou bidimensional se tornam ordenadas entre si, de tal forma que um conjunto de coordenadas é criado para representar diferentes características dos padrões de entrada.

Haykin (2001, p.483) caracteriza uma rede Som como:

... uma formação de um mapa topográfico dos padrões de entrada no qual as localizações espaciais (i.e, coordenadas) dos neurônios na grade são indicativas das características estatísticas intrínsecas contidas nos padrões de entrada, daí o nome “mapa auto-organizável”.

A seguir, o modelo de mapa proposto por Kohonen:

Figura 1 - O Modelo de Kohonen



Fonte: [5] Haykin

O processo de aprendizagem (atualização dos pesos w_{ji}) deste modelo de rede neural aqui estudado acontece em duas fases. A primeira trata-se da Fase de Ordenação, onde a ordenação topológica dos vetores de peso w_{ji} ocorre. Alguns parâmetros importantes nesta fase são a taxa de aprendizagem $h(n)$, adaptativa decrescente a cada atualização de n atualizações, e a função de vizinhança $\phi(h, n, j, i)$. A função de vizinhança representa um papel importante em mapas auto-organizáveis. Especificamente na fase ordenação, permite-se que os neurônios pertencente a uma certa vizinhança do neurônio vencedor i sejam atualizados. Durante esta fase, a vizinhança se reduzirá a 1 (um) par de neurônios ao redor do neurônio vencedor i , ao final de cada atualização.

Já na Fase de Convergência, ou Ajuste Fino, recomenda-se que o número de iterações seja de no mínimo de 500 vezes os números de neurônios na grade. Nesta fase os pesos sofrem um processo de convergência permitindo uma quantização estatística precisa do espaço de entrada. A vizinhança ao redor de um neurônio deve ser mantida igual a 1. Em alguns casos permite-se que ela decresça até 0, desta forma, somente os pesos associados ao neurônio vencedor seriam atualizados. Quanto ao valor de $h(n)$, deve ser bem pequeno e fixo, digamos 0,01, pois o que buscamos é uma precisão do espaço de padrões representados pelos neurônios.

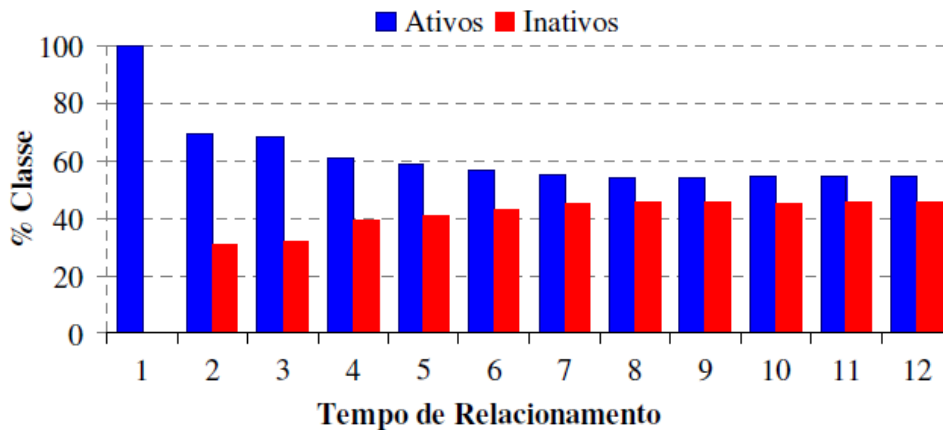
Recomendamos para mais detalhes a leitura de [5]. Seguiremos com a descrição do problema para aplicação dos modelos neurais de Kohonen.

A EMPRESA E O PROBLEMA DE DIFERENCIAÇÃO NO MARKETING

A empresa em estudo trata-se de uma administradora de investimentos com grande atuação na Bovespa (Bolsa de Valores do Estado de São Paulo). Clientes que optam por investir pelo sistema Home Broker executam ordens online, diversificando seus investimentos em até 4 categorias de investimentos: (1) compra e venda de ações; (2) cotas em fundos de investimentos; (3) bolsa de mercadorias e valores futuros (ou Bm&f); (4) e títulos do tesouro direto.

O problema da empresa apresenta-se da seguinte forma: clientes após a inclusão na base de dados e início das operações no sistema home broker, apresentam um decréscimo significativo na atividade até o 4º mês de relacionamento, identificado pela não intenção de continuar investindo. É fato para a empresa que após o 4º mês de relacionamento clientes que decidem por continuar suas operações na bolsa ou outro tipo de investimento o fazem continuamente ao longo de um horizonte de 12 meses ou mais.

Para que isto fique claro, apresentaremos o gráfico a seguir. Para tanto, separamos 12 safras mensais de entrada de cliente ao longo do ano de 2006, e verificou-se o status do cliente, mês a mês, por um período seguinte de 12 meses. Quando no mês de análise, após a inclusão no cadastro da empresa, não era verificada nenhuma operação no sistema home broker, o cliente era marcado como inativo. Contrariamente, recebia a marcação de ativo, caso viesse a efetuar alguma ordem de investimento no mês.

Gráfico 1 - Curvas de Atividade e Inatividade

Fonte: Própria Autoria

A empresa busca reforço no marketing para diminuição das taxas de inatividade após o 4º mês, mas não tem obtido grandes sucessos. Clientes não se identificam com produtos ou serviços lhes ofertado, e a não intenção em continuar operando na bolsa é impulsionada pelo aumento do atrito com a empresa. Para dificultar, clientes não obtêm o retorno esperado no mercado em que operam, e atribuem à empresa esta responsabilidade, por ter-lhes indicado aquela composição de carteira de investimento. Nestes ambientes, acredita-se que boas partes destes problemas vivenciados poderiam ser solucionadas se gestores de relacionamento ao cliente incorporassem em seus processos *estratégias por cluster de acordo com o mercado em que operam*.

A Bovespa disponibiliza para qualquer pessoa física a classificação setorial das empresas e fundos negociados. Tal classificação tem como principal objetivo agrupar as empresas que atuem em estágios similares da cadeia produtiva, ou com produtos/serviços relacionados, e tendam a responder de forma semelhante às condições econômicas. São elas:

Tabela 1 - Classificação Setorial de Fundos e Ações de Empresas Negociadas na Bovespa

Classificação
Bens Industriais
Construção e Transporte
Consumo Cíclico
Consumo Não Cíclico
Financeiro e Outros
Materiais Básicos
Petróleo e Gás
Telecomunicações
TI
Utilidade Pública

Fonte: www.bovespa.com.br

Desta forma, acredita-se que clientes podem ser clusterizados com base na *distribuição de ordens executadas em cada um dos setores apresentados*. Isto servirá de motivação para aplicação das Redes SOM neste artigo apresentado.

Para tanto, utilizaremos uma base de dados composta por uma amostra de clientes novos ao longo de um período de 12 meses, o que nos dá um total de total de 14.799. Variáveis (colunas) são representadas valor da distribuição percentual em cada um dos 10 setores. A seguir, resultados obtidos pela modelagem por Redes SOM.

MODELAGEM POR REDES SOM

Para as 10 variáveis setoriais pertencentes a camada de entrada da rede, foram testadas duas forma de normalização de variáveis. Na primeira, utilizamos uma transformação linear para o intervalo [0;1]. Num segundo momento, uma transformação para variância unitária.

Testamos também as variáveis na forma com foram apresentadas em 3, sem alguma normalização. Sobre isto, conclui-se que a normalização das vaiáveis não tem grande impacto nos resultados da rede, uma vez que os valores da entrada originalmente já estão no intervalo de 0 a 1.

Muitas configurações de mapas foram testadas em conjunto com as normalizações apresentadas anteriormente, sempre da forma quadrática ou retangular. A exemplo: 10x10, 5x10, 15x10, 15x15 e 8x13. Optamos pelo mapa de dimensão 13x8 neurônios pela nítida visualização e interpretação de grupos. Neste último mapa em específico, parâmetros foram ajustados conforme detalhamento a seguir.

Para a função de vizinhança em $N_c(0)$, definimos como 4 (metade de neurônios do lado menor do retângulo). Na fase de ordenação, $N_c(t)$ diminuía linearmente proporcional a t , até o valor 1, por um ciclo de $t=1000$ iterações. Sobre o parâmetro alfa, definimos como 0,5, decrescendo da mesma forma proporcional a t até 0,01 na próxima fase.

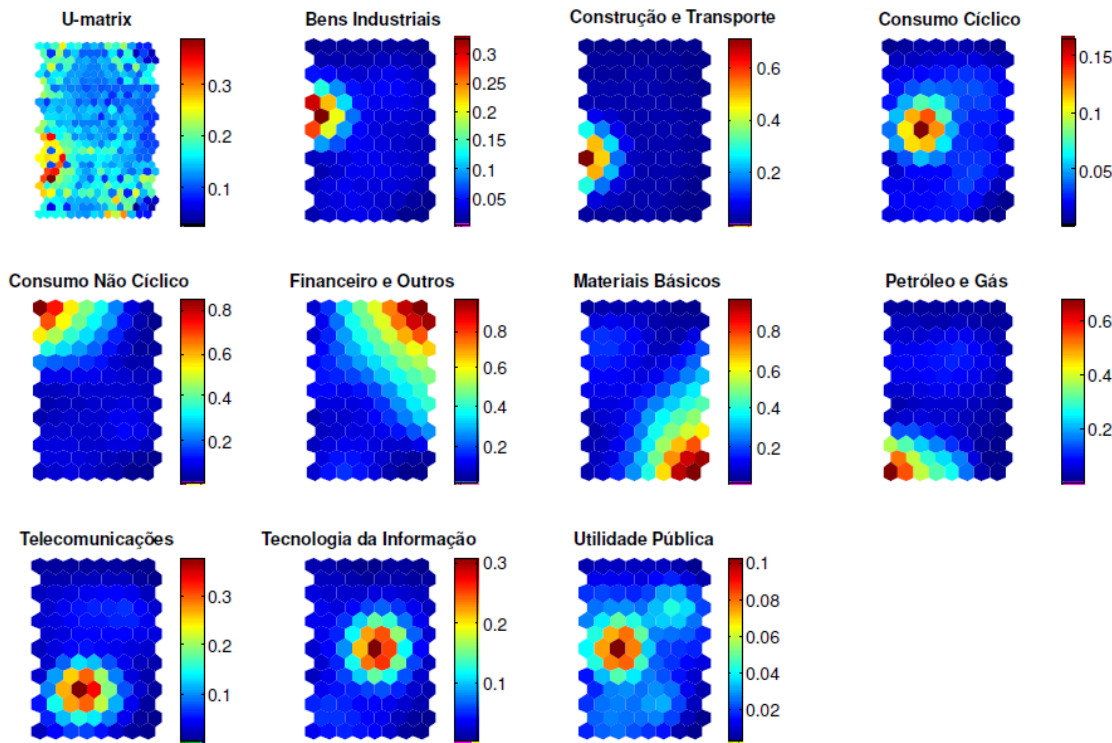
Já na fase de ajustamento fino, com o parâmetro alfa fixo de 0,01, e $N_c(t)$ fixo igual a 1, a atualização dos pesos aconteceu por um ciclo de 50.000 iterações. É importante dizer que o software utilizado para ajuste da rede foi o Matlab 7.0, e que o tipo de aprendizado implementado no Toolbox Som é de aprendizado em Batch, ou seja, os parâmetros são ajustados após a apresentação de todos os padrões a rede.

A matriz de U , o número de clientes representados por cada neurônio, quantidade de clusters considerados, bem como os centróides destes serão apresentados a seguir:

RESULTADOS OBTIDOS

A matriz de U (ou matriz de distância unificada) permite uma visualização unificada da distância entre os neurônios para todas as variáveis de análise. Quanto mais marrom a cor do neurônio, maior é a distancia entre o neurônio vizinho mais próximo. Entre outras palavras, maior é a diferença de perfil do cliente ativado pelo neurônio vizinho mais próximo. Por outro lado, quanto mais azul a cor do neurônio, menor é a distância entre o neurônio vizinho mais próximo, ou seja, há uma maior semelhança de perfil entre o cliente ativado pelo neurônio vizinho.

Figura 2 - Matrizes de Distâncias



Fonte: Software Matlab

Percebe-se que clientes ativados pelos neurônios em cada um dos 4 cantos do mapa representam individualmente o centróide de um cluster, dados por maiores concentrações de ordens executadas respectivamente nos mercados de *Consumo Não Cíclico*, *Financeiro e Outros*, *Materiais Básicos*, e *Petróleo e Gás*. A escala de concentração de cada uma destas variáveis nos mostra que: clientes ativados pelo centróide do Cluster Consumo não Cíclico, executam aproximadamente 80% de suas ordens neste mercado; para o cluster do setor Financeiro a concentração é maior que 80%; da mesma forma acontece para o cluster de Materiais Básicos, com concentração maior que 80%; e para o cluster de Petróleo e Gás, concentração de ordens pouco maior que 60%.

Setores de Tecnologia da Informação, Consumo Cíclico, Bens Industriais, e Utilidade Pública, apresentaram centróides bem próximos. Analisando a região da matriz de U a qual estes neurônios são mais ativados por cada uma das variáveis, percebe-se uma zona azulada característica da pequena distância de perfis de clientes ativados nesta região. Por estes motivos, acredita-se que esta região sozinha representaria um *quinto cluster distinto de clientes*.

Da mesma forma, Telecomunicações e Construção e Transporte apresentaram centroides relativamente próximos, e por isso, decidiu-se agrupar clientes ativados nestas regiões para formarem um *sexto cluster de clientes*. Clientes os quais apresentariam uma grande atividade de ordens executadas nestes mercados.

A figura abaixo mostra o número de clientes ativados por cada um dos neurônios, bem como o agrupamento considerado para formação de cada um dos 6 cluster.

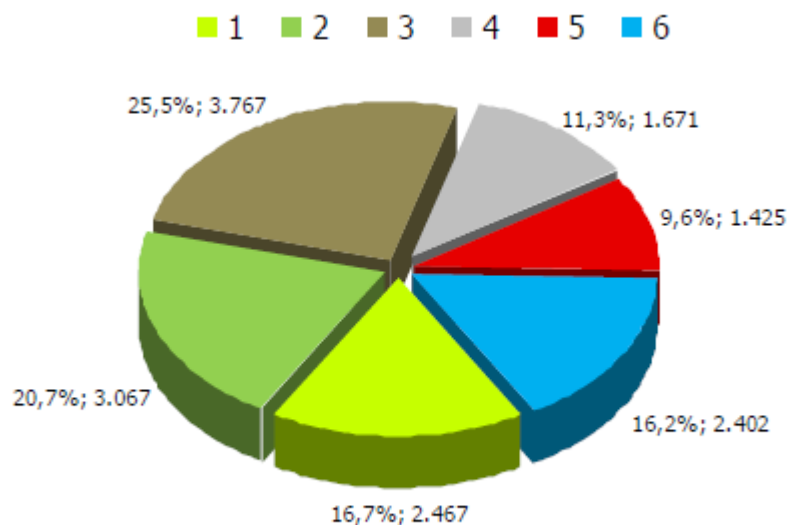
Figura 3 - Quantidade de Clientes Ativados por Neurônio e Formação de Clusters

(i,j)	1	2	3	4	5	6	7	8
1	530	66	150	282	168	132	105	839
2	98	81	107	100	114	33	39	55
3	192	111	190	153	134	90	161	125
4	151	108	148	120	129	180	50	240
5	101	93	97	116	111	108	103	111
6	354	55	87	68	136	102	113	167
7	44	28	226	50	218	78	125	104
8	124	85	158	74	105	113	131	304
9	388	45	39	41	48	80	150	115
10	44	37	233	143	101	95	73	176
11	69	58	135	44	60	98	126	48
12	130	142	123	104	73	90	26	19
13	312	149	217	225	241	103	9	1.320

Fonte: Própria Autoria

A distribuição de clientes por cada um dos clusters pode ser obtida abaixo.

Gráfico 2 - Distribuição de Clientes por Clusters



Fonte: Própria Autoria

Para a caracterização dos centróides finais foi verificado a média de concentração de ordens executadas em cada um dos 10 setores, por clientes pertencentes a cada um dos 6 clusters. A saber:

Tabela 2 - Caracterização dos Centróides Finais

Setor	Cluster - % Médios					
	1	2	3	4	5	6
Bens Industriais	1,8%	1,9%	10,3%	2,5%	2,9%	0,9%
Construção Transporte	1,8%	1,2%	2,9%	32,1%	3,3%	1,1%
Consumo Cíclico	1,2%	1,6%	5,5%	1,6%	2,2%	1,1%
Consumo Não Cíclico	54,4%	7,0%	8,2%	5,1%	7,3%	2,3%
Financeiro e Outros	21,6%	72,0%	23,1%	8,3%	13,2%	4,2%
Materiais Básicos	11,3%	8,7%	20,9%	14,5%	11,1%	82,1%
Petróleo e Gás	3,5%	3,3%	6,3%	6,8%	48,2%	5,4%
Telecom	2,1%	2,6%	3,6%	22,7%	4,4%	1,4%
Tecnologia Informação	1,7%	1,9%	11,5%	3,2%	4,5%	1,2%
Utilidade Pública	1,2%	1,6%	4,1%	2,0%	2,7%	0,7%

Fonte: Própria Autoria

A exceção dos clusters 1, 2, 5 e 6 que apresentaram predominância de concentração de ordens executadas em apenas 1 (um) setor em específico, os clusters 3 e 4, onde houve o agrupamento de 2 ou mais centróides de variáveis, apresentaram concentrações bastante diversificadas. Cluster 3, a exemplo, obteve uma maior concentração em Materiais Básicos, e Financeiro e Outros, enquanto que cluster 4 alta concentração em Construção e Transportes, e Telecom.

CONCLUSÕES E PRÓXIMOS PASSOS

Como observado no tópico 3, a identificação de clusters será útil na diferenciação de ações de marketing. Gestores de relacionamento ao cliente poderão a partir do 4º mês de relacionamento do cliente fazer ofertas personalizadas de produtos e serviços, maximizando assim o retorno de ações. Com isso, esperam-se diminuir as taxas de evasão de clientes a partir do 5º mês de relacionamento.

É importante dizer que a aplicação aqui feita através de Mapas Auto-organizáveis teve como objetivo a exploração do potencial de aplicação desta técnica num ambiente empresarial, e não somente a resolução em específico do problema da empresa aqui abordada.

O raciocínio aqui desenvolvido poderia ser facilmente adaptado para empresas nos mais diversos ramos de atividade.

No entanto, alguns pontos de estudo são necessários para a melhoria e aperfeiçoamento da técnica utilizada. Como exemplo, na formação de clusters conforme figura 3, é de grande dúvida para o analista a determinação das fronteiras entre clusters. Por estes motivos, clientes, mesmo que se encontre em zonas de fronteira entre dois ou mais Clusters, onde não há a predominância de um perfil em específico, podem ser classificados em grupos com menor semelhança. Para que isso não ocorra, alguma heurística adicional deve ser considerada, como maximização da variância entre clusters. Uma alternativa seria a aplicação de alguma técnica de clusterização nos próprios clusters ao invés, conforme proposto por Vesanto em [7].

REFERÊNCIA BIBLIOGRÁFICAS

- REINARTZ, Werner. KUMAR, V. **Customer Relationship Management, A Database Approach**. John Wiley & Sons, June 2005.
- HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles of Data Mining**. The MIT Press, Cambridge, 2001.
- BERRY, Michael; LINOFF, Gordon. **Data Mining Techniques for Marketing, Sales, and Customer Relationship Management**. John Wiley & Sons, 2004.
- TWO CROSS CORPORATION. **Introduction to Data Mining and Knowledge Discovery**. 3rd ed.: USA, 1999.
- HAYKIN, Simon. **Redes Neurais, Princípios e Práticas**. Paulo Matins Engel. 2ed. Porto Alegre: Bookman, 2001.
- LAROSE, Daniel. **Data Mining Methods and Models**. Central Connecticut State University: John Wiley & Sons, 2006.
- VESANTO, Juha; ALHONIEMI, Esa. **Clustering of the Self-Organizing Map**. IEEE Transactions on Neural Networks, Vol. 11, No 3, Maio de 2000.
- MARQUES, Márcio Candeias. **Comparação entre os métodos de agrupamentos Kmeans e Mapa de Kohonen (SOM) em análise de pesquisa de mercado**. Revista Inteligência Computacional Aplicada da Puc-Rio: Junho de 2008.
- HAIR, Joseph, *et al.* **Análise Multivariada de Dados**. Porto Alegre: Bookman, 2005.