Ranking Model Applying Self-Organizing Maps and Factor Analysis

D. Steffen and A. Chaves Neto

***Abstract*—The present work aims to use Self-Organizing Maps (SOM) to group ranking results obtained by factor analysis. In this paper we propose the application of Factor Analysis at a stage prior to the application of the SOM, in order to use the factor scores of latent variables as input data for network training, as these scores carry practically all information regarding the variables of the lattice problem, losing only an insignificant portion of variability. Thus, the resulting map will present the cluster according to the efficiency shown in the Factor Analysis. The application refers to the performance evaluation of 50 employees of a company by means of scores on four psychological assessment tests (exams) and three indicators of sales results. The result of ranking these vendors according to performance was grouped into four homogeneous clusters.**

***Index Terms*—Factor Analysis, Ranking Model, Self-Organizing Maps.**

I. Introdução

Quando se pretende classificar atividades do sistema produtivo destacando unidades com bom desempenho, técnicas de Análise Multivariada tais como a Análise Fatorial juntamente com a Análise Envoltória de Dados (DEA – Data Envelopment Analysis) apresentam-se muito citadas na literatura. A Análise Fatorial é uma técnica comprovadamente bem-sucedida, isto pode ser constatado nos artigos encontrados em [1], [2] e [3].

Em economia, por exemplo, a concessão de crédito a empreendedores é uma prática realizada por muitas organizações. A decisão dessa concessão é fundamentada em técnicas de classificação, ou seja, de rotulação do cliente como "bom pagador" ou "mal pagador". A análise fatorial pode ser empregada nestes casos para identificar as correspondências entre as variáveis, sendo possível ranquear os consumidores em relação à sua capacidade de pagamento.

Este trabalho tem por finalidade apresentar a técnica de Redes Neurais conhecida por Mapas Auto-Organizáveis de Kohonen (SOM – *Self-Organizing Maps*), no agrupamento de itens com características numéricas em clusters homogêneos a partir do ranqueamento obtido por Análise Fatorial. Isto é feito de forma que os resultados do ranqueamento situem-se em grupos de características similares.

D. Steffen é professor Titular B da Universidade Comunitária da Região de Chapecó, Chapecó/SC Brasil. daniel_steffen@unochapeco.edu.br.

A. Chaves Neto é professor sênior no Programa de Pós Graduação em Métodos Numéricos em Engenharia da Universidade Federal do Paraná, Curitiba/PR, Brasil. anselmo@ufpr.br.

A metodologia proposta apresenta um procedimento de discriminação dos itens juntando em grupos que poderão ser classificados como "ótimos", "bons", "regulares" e "ruins", por exemplo, ou em outra categorização.

A rede SOM é um algoritmo de Redes Neurais muito utilizado em aprendizagem não supervisionada, agrupamento, classificação e visualização de dados. Várias publicações estão relatadas na literatura e vários projetos comerciais empregam o SOM como um procedimento para resolver problemas considerados difíceis do mundo real [4]. Sistemas híbridos envolvendo essa rede neural também foram aplicados com sucesso em estudos tais como em [5], no qual se aplica a técnica DEA, juntamente com uma Rede SOM, produzindo um novo ranqueamento em um sistema modelado para ser imparcial e considerado como "justo" pelos grupos envolvidos na análise.

Nesta proposta, a novidade é a aplicação da Análise Fatorial em uma etapa anterior à aplicação do SOM. Para que a eficiência dos indivíduos verificada na Análise Fatorial seja refletida na aplicação do SOM, utiliza-se os escores fatoriais das variáveis latentes como dado de entrada para o treinamento da rede, pois esses escores carregam praticamente toda a informação referente às variáveis do problema, perdendo apenas uma parcela insignificante da variabilidade.

Esta aplicação em conjunto, Análise Fatorial e SOM, se faz necessária para a posterior classificação dos clusters, pois após a Análise Fatorial se torna conhecida a importância das variáveis latentes pelo grau de explicação que cada uma delas carrega. Com isso, é possível ranquear o cluster como eficiente ou não eficiente, podendo inclusive identificar as características necessárias a cada indivíduo para ser bem classificado.

A mesma metodologia é aplicada utilizando o K-means para fins de comparação com o SOM.

II. Análise Fatorial

A Análise Fatorial é uma técnica multivariada proposta primeiramente no início do século 20 por Charles Spearman [6] para estudar problemas relacionados à psicologia educacional e, particularmente, ao estudo de variáveis latentes como a inteligência. Spearman estudou a hipótese de que diferentes testes de habilidade mental poderiam ser explicados por um fator comum, a inteligência. Thurstone [7] desenvolveu a ideia de "multiple factors analysis". Hotelling [8] propôs o método das componentes principais que permite o cálculo da única matriz de fatores ortogonais. De acordo com [9], a Análise Fatorial é a principal e a mais antiga técnica de análise multivariada.

De acordo com Kubrusly [10], ao contrário do modelo de

Análise das Componentes Principais, o modelo de Análise Fatorial procura reproduzir da melhor forma possível a correlação entre as variáveis originais. Seu principal propósito é descrever a estrutura de covariâncias entre as variáveis em termos em termos de um número menor de variáveis não observáveis (latentes) chamadas fatores. Resumindo, as variáveis são agrupadas levando-se em conta suas correlações. Portanto, as variáveis dentro de cada grupo (fator) terão forte correlação. Em outras palavras, a Análise Fatorial explica as correlações entre um conjunto grande de variáveis em termos de um conjunto com poucas variáveis, mas que estão nas direções de maior variabilidade. No livro do Hair [11], os autores descrevem os seguintes passos para a realização de uma análise fatorial: formulação do problema; construção da matriz de correlação; escolha do método de estimação dos fatores; interpretação dos fatores; cálculo das cargas fatoriais e ajuste do modelo.

Assim, considerando os objetivos deste trabalho, tem-se inicialmente a matriz de dados originais para a qual se calcula a matriz de correlação R e, então, estimam-se os pesos dos fatores por componentes principais.

De acordo com Johnson e Wichern [12], seja o vetor aleatório $X' = [X_1, X_2, \dots, X_p]$, com p variáveis e com $E(\underline{X}) = \underline{\mu}' = [\mu_1, \mu_2, \dots, \mu_p]$ e matriz de covariância $V(\underline{X}) = \Sigma_{p\times p}$. O vetor $\underline{X}$ é linearmente dependente sobre variáveis não observáveis (latentes) $F_1, F_2 \dots, F_m$, $m<p$ chamadas fatores comuns e p fontes de variação aditivas, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, chamadas de erros ou fatores específicos. Cada vetor observado e padronizado pode ser escrito no modelo fatorial como em (1):

$$\underline{Z}_{p\times 1} = L_{p\times m}.\underline{F}_{m\times 1} + \underline{\varepsilon}_{p\times 1} \quad (1)$$

onde L é a matriz de pesos ou carregamentos de ordem $p \times m$, $\underline{F}$ é o vetor de fatores extraídos de dimensão ***m,*** e $\underline{\varepsilon}$ é o vetor de erros ou fatores específicos de dimensão *p*.

A normalidade dos dados é necessária somente quando um teste estatístico é aplicado para revelar a significância dos fatores. Mas, é necessário medir o quanto uma variável pode ser explicada por outra variável (multicolinearidade), ou seja, a matriz de correlação do vetor $\underline{X}$ é significativamente diferente da matriz identidade. Um modo de verificar a adequação dos dados à Análise Fatorial é aquele feito pelo critério de Kaiser-Meyer-Olkin (KMO).

De acordo com Marroco, J. [13], o critério KMO é uma medida de homogeneidade das variáveis que compara as correlações simples com as correlações parciais observadas entre as variáveis. Em outras palavras, o critério KMO identifica se o modelo de Análise Fatorial está adequadamente ajustado aos dados. Essa medida é dada por (2):

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (2)$$

onde r_{ij} é o coeficiente de correlação linear simples entre X_i e X_j; e a_{ij} é o coeficiente de correlação parcial entre X_i e X_j.

Os valores da estatística KMO variam de 0 a 1 e avaliam a adequação da amostra quanto ao grau de correlação parcial entre as variáveis. Sugere-se que valores de KMO<0.5 são indicativos de que a matriz de correlação não se presta a análise fatorial.

Para a interpretação dos resultados da Análise Fatorial utiliza-se um procedimento de rotação de fatores. O método de rotação mais utilizado é a "Rotação Ortogonal", que pode ser executada pelo método Varimax e na qual os eixos são mantidos a 90º, consequentemente os fatores são não correlacionados. O critério Varimax define, $\tilde{e}_{ij}^{*} = \hat{e}_{ij}^{*2}/\hat{h}_{ij}$ como sendo os coeficientes finais após a rotação corrigidas pela raiz quadrada das comunalidades. O procedimento varimax seleciona a transformação ortogonal T que maximiza:

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{j=1}^{p}\tilde{e}_{ij}^{*4} - \frac{1}{p}\left(\sum_{i=1}^{p}\tilde{e}_{ij}^{*2}\right)^2\right] \quad (3)$$

Espera-se encontrar grupos de coeficientes grandes e coeficientes desprezíveis em qualquer coluna da matriz de cargas rotacionadas. Após a transformação T ser determinada, os pesos $\tilde{e}_{ij}^{*}$ são multiplicados pelas comunalidades $\hat{h}_{ij}$, de forma que as comunalidades originais sejam preservadas. O método de rotação tem o objetivo de simplificar as linhas e colunas da matriz fatorial para facilitar a interpretação [11].

Os escores fatoriais f_j, $j = 1,2, \dots, n$ podem ser calculados pelo Método dos Mínimos Quadrados Ponderados.

$$\hat{f}_j = \left(\hat{L}'.\hat{\Psi}^{-1}\hat{L}\right)^{-1}.\hat{L}'\hat{\Psi}^{-1}.\left(\underline{x}_j - \bar{\underline{x}}\right) \quad (4)$$

em que $\hat{f}_j$ são os escores fatoriais, $\hat{L}$ a matriz de cargas fatoriais e $\hat{\Psi}$ as variâncias específicas. E, os escores classificatórios são obtidos ponderando-se os escores fatoriais originais pela importância de cada fator, mensurada pelo autovalor. O escore bruto é dado por (5):

$$E_j = \frac{\sum_{i=1}^{k} f_j.\hat{\lambda}_i}{\sum_{i=1}^{k}\hat{\lambda}_i} \quad (5)$$

com $\hat{\lambda}_i$ sendo o autovalor i correspondente ao fator. Um autovalor de grande magnitude corresponde a um peso maior. Assim, a unidade que apresentar escore fatorial de grande valor corresponde a um grande autovalor e que portanto, tende a ter melhor classificação.

III. Rede Som

As redes SOM (*Self-Organizing Maps*) foram desenvolvidas por Teuvo Kohonen [16] na década de 1980, desde então o algoritmo original vem sofrendo modificações para atender a critérios que inicialmente não foram abordados. Os mapas auto-organizáveis de Kohonen, como também é conhecido, fazem parte de um grupo de redes neurais baseadas em modelos de competição em que os neurônios de saída competem entre si para serem ativados [14].

Esse tipo de rede utiliza método de treinamento não

supervisionado. Isso significa que o resultado é formado pelas propriedades inerentes aos dados em si e não se tem conhecimento, a priori, do que é verdadeiro ou falso. Portanto, nenhum conhecimento prévio sobre a estrutura dos dados é necessário para o uso do algoritmo. Assim, a própria rede neural busca encontrar similaridades baseando-se apenas nos padrões de entrada [15].

O modelo de rede proposto por Kohonen [16] tem o objetivo de captar as características essenciais presentes nos dados de entrada e apresenta um resultado correspondendo a um mapa topográfico, com otimização do posicionamento de um número fixo de vetores iniciais. Inicialmente, a proposta é agrupar os dados de entrada que são semelhantes formando classes ou agrupamentos (*clusters*). A localização espacial dos neurônios na grade, após o aprendizado, são indicativa das estatísticas presentes nos padrões de entrada.

Esse tipo de rede neural possui duas camadas, a primeira corresponde a camada de entrada, a segunda aos neurônios, sendo esta camada ordenada, dita competitiva, e que é gradualmente adaptada para responder seletivamente às entradas similares.

O algoritmo está dividido em três etapas: competição, cooperação e adaptação sináptica. Na primeira etapa, dependendo da resposta de cada neurônio, um único elemento é declarado vencedor. Em seguida, são determinados os elementos que compõem a vizinhança do neurônio vencedor para ser associado um coeficiente de cooperação. Com base nesses dados, o neurônio vencedor recebe um incremento para correção da sua função de ativação e, conforme o fator de vizinhança, um percentual do incremento recebido pelo neurônio vencedor é aplicado aos neurônios vizinhos [17].

A Fig. 1 ilustra a estrutura básica de uma rede SOM, apresentando as duas camadas: a camada de entrada e a camada de saída, esta última, representada por um grid bidimensional.

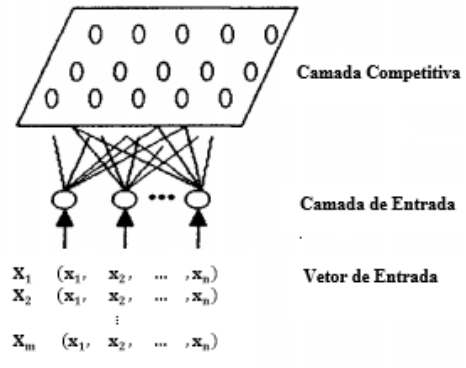

Fig. 1. Estrutura básica da rede de Kohonen.

Os passos são os seguintes, conforme descrito por [18]:

Passo 1: Iniciar os pesos w_{ij} aleatoriamente de todas as unidades na camada competitiva.

Passo 2: Aplicar um vetor de entrada $\underline{X}$ do conjunto de treinamento e determinar o neurônio vencedor *k*. A métrica utilizada é a distância euclidiana.

$$d(w_i, x_k) = \sqrt{\sum_{j=1}^{p}\left[x_{kj}(t) - w_{ij}(t)\right]^2} \quad (6)$$

O neurônio vencedor *c*, é escolhido de acordo com (7):

$$\|x - w_c\| = \min_i\{\|x - w_i\|\} \quad (7)$$

onde $\|\cdot\|$ denota a distância obtida.

Passo 3: Atualizar os pesos sinápticos da unidade vencedora e de seus vizinhos usando (8):

$$w_i(t+1) = w_i(t) + \alpha(t).g(i,k).\left(x(t) - w_i(t)\right) \quad (8)$$

com $\alpha(t)$ sendo a taxa de aprendizagem e $g(i,k)$ função de vizinhança que decresce com a distância entre as unidades *i* e *k*.

Uma função interessante a ser utilizada como função de vizinhança é a Gaussiana, dada por (9):

$$g(i,k) = exp\left(-\frac{\|r_i - r_k\|}{2\sigma^2(t)}\right) \quad (9)$$

em que o termo $\|r_i - r_k\|$ é a distância entre o neurônio "*i*" vencedor e o neurônio "*k*" que está sendo atualizado. O parâmetro σ define a largura da função que deve ser decrescente no tempo. Em geral, é utilizada a função exponencial (10):

$$\sigma(t) = \sigma(0).exp\left(-\frac{1}{\tau_1}\right) \quad (10)$$

com $\sigma(0)$ sendo o valor inicial para σ; τ_1: constante de tempo do SOM. Para que a taxa de aprendizagem nunca caia para zero utiliza-se $\tau_1 = \frac{NIter}{\log\sigma(0)}$, onde *NIter* é o número de iterações.

Passo 4: Atualizar a função de vizinhança g e a taxa de aprendizagem σ.

Passo 5: Repetir a partir do passo 2 para cada um dos vetores do conjunto de treinamento até atingir o número de épocas estimadas ou até que os pesos se estabilizem.

IV. Aplicação

A. Dados Utilizados

No desenvolvimento deste trabalho foi utilizado um banco de dados referente a informações sobre 50 vendedores de uma empresa, este caso foi adaptado do encontrado em [12]. A empresa avaliou o desempenho de 50 funcionários por meio de escores em quatro testes (exames) de avaliação psicológica e três indicadores de resultados em vendas. Os indicadores de vendas foram: crescimento das vendas, rentabilidade das vendas e vendas de novas contas (novos clientes). As medidas desses indicadores foram convertidas para uma escala em que 100 indica o desempenho médio. E, também, cada um dos vendedores foi submetido a quatro testes com o propósito de medir a criatividade, raciocínio mecânico, raciocínio abstrato e habilidade em matemática. As variáveis foram denotadas da seguinte forma:

X_1: Crescimento das vendas;

X_2: Rentabilidade das vendas;
X_3: Vendas para novos clientes;
X_4: Criatividade;
X_5: Raciocínio mecânico;
X_6: Raciocínio abstrato;
X_7: Habilidade em matemática.

V. Resultados da Metodologia Proposta

A. Análise Fatorial

A aplicação da Análise Fatorial corresponde a primeira etapa da metodologia proposta em que são determinados os escores fatoriais para posterior aplicação do SOM.

A Análise Fatorial é aplicada sob os dados de entrada da matriz X de ordem $n \times p$ e se calcula a matriz de correlação R de ordem $p \times p$ do vetor aleatório observado em n=50 observações. Para a utilização da Análise Fatorial é importante que as premissas para sua aplicação sejam observadas. Assim, aplicou-se o critério KMO. A estatística KMO mede o grau de adequação da amostra quanto ao grau de correlação parcial entre as variáveis, resultando um valor de 0,67. Sendo considerada boa a adequação dos dados à Análise Fatorial.

Após a obtenção da matriz de correlação R calcula-se os p pares de autovalor e autovetor dessa matriz. Para determinar o número de fatores (variáveis latentes) da Análise Fatorial utilizou-se nesta aplicação o critério da percentagem da variância total explicada. Utiliza-se um número de fatores que explique o máximo da variabilidade dos dados, nesta aplicação, utiliza-se um número m de fatores que explique ao menos 90% dessa variabilidade.

Seguindo este critério, do banco de dados original foram extraídos 3 fatores que representam 92,3% da variabilidade. Na Tabela I são apresentados os carregamentos (pesos) dos fatores rotacionados pelo critério Varimax.

TABELA I
CARREGAMENTOS FATORIAIS ROTACIONADOS

Variáveis	Fator 1	Fator 2	Fator 3
X_1: Crescimento das Vendas	0,779446	0,387000	0,451724
X_2: Rentabilidade das Vendas	0,908163	0,356104	0,189007
X_3: Vendas de Novas Contas	0,616270	0,547521	0,483554
X_4: Criatividade	0,212867	0,952391	0,047236
X_5: Raciocínio Mecânico	0,552345	0,607037	0,145567
X_6: Raciocínio Abstrato	0,286483	0,060367	0,949734
X_7: Habilidade em Matemática	0,909293	0,180566	0,327866
Variância Explicada (%)	**71,92**	**13,33**	**7,11**

O Fator 1 é fortemente correlacionado com a habilidade em matemática e rentabilidade de vendas com pesos 0,909 e 0,908, respectivamente. Já o Fator 2 é fortemente influenciado pela criatividade com peso 0,952 e o Fator 3 pelo raciocínio abstrato com peso 0,949. Os fatores obtidos pela Análise Fatorial são descritos a seguir:

Fator 1: as variáveis X_1, X_2, X_3 e X_7 são as variáveis de maior peso para o Fator 1 e foi aqui interpretado como "indicadores de vendas", este fator explica a maior parte de variabilidade, 71,92% da variabilidade total.

Fator 2: as variáveis X_4 e X_5 são interpretadas como "resultados nos testes", este fator explica 13,33% da variabilidade total.

Fator 3: a variável X_6 é interpretada como "abstração de resultados". Este fator explica 7,11% da variabilidade total.

Na Tabela II, são apresentadas as comunalidades e variâncias específicas, que por sua vez revelam que todas as variáveis são importantes no estudo, haja vista que suas comunalidades são bem altas, consequentemente as variâncias específicas são bem baixas. Assim, o percentual da variância total atribuído aos 3 fatores comuns é grande e, consequentemente, o erro devido à aleatoriedade é pequeno. Isto é imprescindível para a verossimilhança da classificação [2].

TABELA II
COMUNALIDADE E VARIÂNCIA ESPECÍFICA

Variáveis	Comunalidade h_i^2	Variância Específica Ψ_i
X_1	0,9613	0,038
X_2	0,9872	0,012
X_3	0,9133	0,086
X_4	0,9545	0,045
X_5	0,6947	0,305
X_6	0,9877	0,012
X_7	0,9669	0,033

Na sequência foram obtidos os coeficientes dos escores fatoriais estimados a partir dos pesos fatoriais rotacionados para as n=50 observações (vendedores). Esses resultados estão na Tabela III e os valores $f_j, j = 1,2,\ldots,n$ são calculados pelo Método dos Mínimos Quadrados Ponderados (Eq. 4).

TABELA III
ESCORES FATORIAIS ROTACIONADOS

Vend.	Fator 1	Fator 2	Fator 3	Vend.	Fator 1	Fator 2	Fator 3
1	-3,75	-2,40	-2,18	26	-2,64	-1,40	-2,71
2	-5,43	-3,78	-2,44	27	3,89	2,53	2,20
3	-2,53	-2,14	-1,68	28	4,25	1,49	1,80
4	0,72	0,91	1,13	29	-4,13	-2,92	-0,53
5	0,92	0,16	0,96	30	3,93	1,32	1,63
6	-2,39	-1,35	-0,81	31	4,48	3,71	1,75
7	-2,58	-1,92	-1,69	32	-5,61	-2,29	-3,39
8	8,02	5,79	5,26	33	-1,17	-1,44	-0,37
9	1,52	0,68	1,60	34	-3,35	-1,96	-1,14
10	3,61	2,34	1,34	35	4,96	4,15	1,87
11	1,80	1,25	1,29	36	3,74	0,60	3,04
12	0,52	0,26	-1,06	37	-1,96	1,15	-1,82
13	2,71	2,68	1,33	38	-0,16	0,08	1,51
14	-0,52	-1,50	0,07	39	5,30	3,72	2,70
15	-0,40	-0,24	-1,08	40	4,39	3,19	2,27
16	-7,04	-4,62	-5,28	41	1,61	0,48	1,62
17	0,02	-0,31	0,29	42	-2,46	-0,55	-3,00
18	1,40	0,47	0,77	43	2,96	3,38	0,92
19	-0,57	-1,70	0,76	44	-7,76	-6,66	-3,65
20	1,00	2,25	0,67	45	-1,62	-1,50	-0,60
21	-5,20	-2,79	-3,71	46	3,01	2,95	2,10
22	1,25	-1,35	0,70	47	-4,05	-2,63	-1,64
23	-6,01	-3,70	-3,93	48	-7,48	-4,65	-4,16
24	2,66	1,38	1,70	49	1,84	1,24	1,54
25	5,00	3,87	2,59	50	3,30	1,73	1,45

A seguir, calcula-se os escores classificatórios ponderando os escores fatoriais originais pela importância de cada fator, representada pelo seu respectivo autovalor. Quanto maior o valor do escore classificatório, melhor posicionado no ranking

estará o indivíduo. Esses escores estão ordenados de forma decrescente (Esc.) na Tabela IV. Com essa ordenação gera-se o ranking do vendedor mais eficiente para o menos eficiente (Rank. Vend.). Na classificação obtida, o 8º vendedor é o mais eficiente com escore 10 e o 44º é o menos eficiente apresentando escore 0,003.

TABELA IV
RESULTADO DA CLASSIFICAÇÃO POR ANÁLISE FATORIAL

Rank (Vend)	Esc.	Rank (Vend)	Esc.	Rank (Vend)	Esc.	Rank (Vend)	Esc.
8	10,00	**43**	6,873	**38**	4,934	**34**	2,911
39	8,232	**13**	6,692	**17**	4,931	**1**	2,606
25	8,079	**24**	6,561	**15**	4,637	**47**	2,450
35	8,048	**49**	6,106	**14**	4,511	**29**	2,439
31	7,750	**11**	6,072	**19**	4,502	**21**	1,725
40	7,678	**41**	5,916	**33**	4,153	**32**	1,573
28	7,414	**9**	5,888	**37**	3,912	**2**	1,569
27	7,344	**18**	5,761	**45**	3,898	**23**	1,195
30	7,217	**20**	5,715	**6**	3,497	**16**	0,494
10	7,134	**22**	5,498	**42**	3,422	**48**	0,318
36	7,120	**5**	5,483	**3**	3,299	**44**	0,003
50	6,920	**4**	5,462	**7**	3,294		
46	6,917	**12**	5,180	**26**	3,258		

.

B. Treinamento da Rede Neural

A rede SOM, aplicada em conjunto com a Análise Fatorial permite a obtenção de clusters homogêneos, tendo como dado de entrada para o treinamento da rede neural os escores fatoriais das variáveis latentes. O propósito da utilização desses dados e não os dados originais, se deve ao fato de que o interesse dessa pesquisa é agrupar indivíduos conforme ranqueamento realizado pela Análise Fatorial e estes escores representam as informações das observações nas direções de maior variabilidade.

A categorização dos clusters torna-se possível, pois as características de cada neurônio/cluster podem ser comparadas com o poder de explicação de cada variável latente, sendo possível, portanto, a distinção dos clusters como "mais eficientes" ou "menos eficientes" adotando os critérios de classificação da Análise Fatorial. A necessidade de conhecer a importância das variáveis no estudo para a classificação do cluster torna o uso das duas técnicas em conjunto necessária.

Este modelo também pode ser aplicado utilizando outros métodos de clusterização mais simples em substituição ao SOM, como o K-means, por exemplo. Apesar do K-means ser uma técnica simples de ser implementada e ter custo computacional baixo, apresenta algumas desvantagens em relação ao SOM, que além de ser suscetível a problemas quando os clusters são de diferentes tamanhos e densidades a quantidade de clusters precisa ser sempre especificada pelo usuário, no caso do SOM a clusterização pode ocorrer de forma automática [19].

Para o estudo de caso apresentado, devido a natureza dos dados, optou-se pela formação de um número fixo de 4 clusters, pois não se observou a necessidade de discriminação em um número maior de clusters. No entanto, em outras aplicações pode ser necessário que o algoritmo investigue a quantidade apropriada de clusters, isso pode ser obtido mais facilmente com o SOM, utilizando uma malha de dimensão maior na execução do algoritmo.

O treinamento da Rede SOM ocorre no ambiente do Software R, o script de execução permite ajustar todos os parâmetros da rede SOM. O Mapa de Kohonen foi desenhado e testado com diversas topologias de malhas bidimensionais para obtenção dos clusters. A categorização pretendida para este conjunto de dados será "ótimo", "bom", "regular" e "ruim", portanto serão necessários 4 clusters. As malhas consideradas para este propósito foram: 1×4, 2×2 e 4×1. Todas as malhas resultam em 4 neurônios. O treinamento com essas três malhas **gerou** resultados muito semelhantes, portanto considerou-se o treinamento com a malha 2×2.

O primeiro passo para a obtenção dos resultados é a normalização dos dados. A normalização é feita em dois passos. Primeiramente se faz a padronização, em que μ é a média e σ é o desvio padrão dos dados obtendo $Z = \frac{X-\mu}{\sigma}$. A partir dos dados padronizados procede-se a segunda transformação, ou seja, colocam-se os dados padronizados no intervalo [0,1]. Para isso, utiliza-se a função Sigmóide $f(x) = \frac{1}{1+e^{-x}}$ com x sendo um número real. A Fig. 2 apresenta o gráfico típico dessa função.

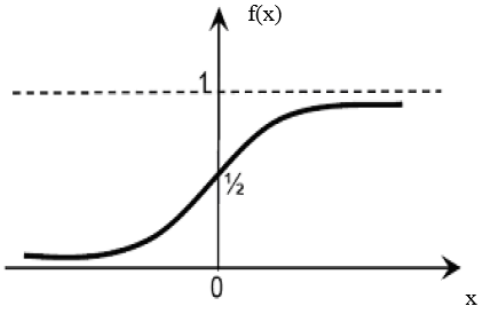

Fig. 2. Gráfico da função Sigmóide. Este tipo de função é utilizada para a padronização de dados, colocando-os entre 0 e 1.

A base de dados para entrada no algoritmo continua uma matriz de ordem 50×3, em que as linhas representam os Vendedores (Vend.) e as colunas as variáveis latentes (fatores).

A topologia 2×2 utilizada neste estudo, gera uma saída de 4 neurônios correspondentes a cada cluster. O motivo de se utilizar 4 neurônios no treinamento se deve à classificação pretendida, neste caso será: "ótimos", "bons", "regulares" e "ruins". Desse modo, agrupando os Vendedores em 4 clusters, estamos juntando os vendedores que possuem as mesmas características e, então, obter uma classificação de acordo com o desempenho. Esses neurônios em conjunto com a matriz de distâncias unificadas (U-mat), que contém as distâncias Euclidianas calculadas para os pares dos neurônios vizinhas, possibilitam a definição dos agrupamentos.

O processo de organização topológica ocorre primeiro onde dados semelhantes ocupam regiões específicas do mapa e, diante disso, são verificadas as distâncias em que neurônios próximos são associados a um mesmo grupo (cluster). A densidade dos dados também pode ser utilizada para definir as fronteiras entre agrupamentos diferentes, principalmente em mapas maiores, nos quais os neurônios de transição entre um grupo e outro ficam vazios, ou seja, não possuem dados associados ao seu modelo [20].

Na fase de treinamento da rede de Kohonen, a estabilidade na atualização dos pesos ocorreu com a execução de 10.000 iterações, a partir disso houve estabilidade nos pesos. A Tabela V apresenta a matriz de pesos sinápticos da rede neural após o treinamento, esses são os "pesos" das variáveis (fatores) que obtiveram maior influência na formação do perfil dos clusters.

TABELA V
MATRIZ DE PESOS

Cluster	Fator 1	Fator 2	Fator 3
V_1	0,7432	0,7208	0,7171
V_2	0,5581	0,5364	0,6048
V_3	0,1630	0,1926	0,1606
V_4	0,3535	0,3716	0,3562

O Fator 1 é constituído por 4 das 7 variáveis originais e explica a maior parte da variabilidade, cerca de 71,92%, seguido pelo Fator 2, com 2 variáveis originais explicando 13,33%, diminuindo o grau de explicação dos fatores até que o último fator explica somente 7,11% da variabilidade total. Portanto, o primeiro fator tem maior importância para obtenção dos índices classificatórios dos clusters, seguido dos demais.

Observa-se na matriz da Tabela V que o Neurônio 1 (V_1) possui o maior peso para o Fator 1, igual a 0,7432, seguido do Neurônio 2 (V_2) com peso 0,5581. Isto pode ser visualizado na Fig. 3, de uma forma melhor.

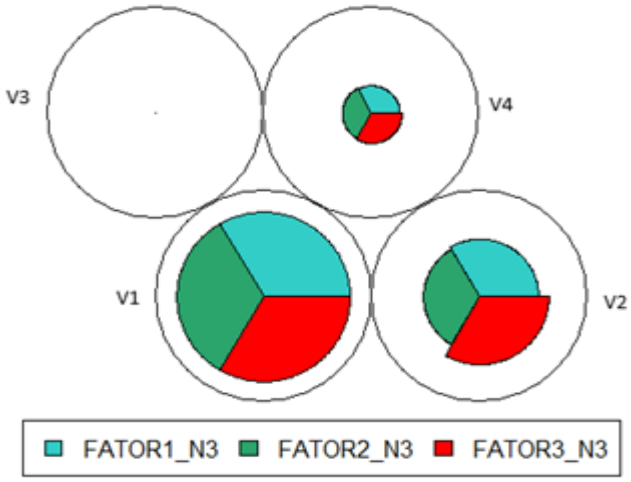

Fig. 3. Vetor de pesos dos Neurônios.

Na Fig. 3, a área do setor circular do peso é diretamente proporcional ao vetor do peso (que está sempre entre 0 e 1) e pode-se perceber que o Neurônio 1 (V_1) possui os maiores pesos para todos os fatores, seguidos intuitivamente pelos neurônios 2, 4 e 3, respectivamente. Para obter uma classificação mais precisa, levando em conta o poder de explicação de cada fator, utiliza-se a expressão (11):

$$\underline{CN} = W^*.\underline{VE} \qquad i = 1,2,3 \tag{11}$$

onde $\underline{CN}$ é o índice classificatório do neurônio/cluster; $\underline{VE}$ é o valor da variância explicada pelo i-ésimo fator obtida na análise fatorial; e W^* é a matriz de pesos (Tabela V). Os resultados estão na Tabela VI.

TABELA VI
RANKING DOS CLUSTERS

Cluster	Pesos	Cluster Rank	Pesos Ordenados
1	0,683953	**1**	0,683953
2	0,520468	**2**	0,520468
3	0,151466	**4**	0,325414
4	0,325414	**3**	0,151466

Quanto maior o valor do "peso", melhor classificado está o Cluster. A classificação dos vendedores pelos escores da análise fatorial é feita ordenando-se os escores em ordem decrescente, logo a ordenação dos clusters fica de acordo com a sequência 1, 2, 4 e 3, conforme exposto na 3ª e 4ª coluna da Tabela VI. Os resultados da classificação estão na Tabela VII.

TABELA VII
RESULTADO DA CLASSIFICAÇÃO COM 4 CLUSTERS (AF/SOM)

Ranking	Número do Vendedor	Cluster SOM	Escore Fatorial	Classificação Cluster
1º	8	1	10,00	
2º	39	1	8,232	
3º	25	1	8,079	
4º	35	1	8,048	
5º	31	1	7,750	
6º	40	1	7,678	
7º	28	1	7,414	
8º	27	1	7,344	**"ótimos"**
9º	30	1	7,217	
10º	10	1	7,134	
11º	36	1	7,120	
12º	50	1	6,920	
13º	46	1	6,917	
14º	43	1	6,873	
15º	13	1	6,692	
16º	24	1	6,561	
17º	49	2	6,106	
18º	11	2	6,072	
19º	41	2	5,916	
20º	9	2	5,888	
21º	18	2	5,761	
22º	20	2	5,715	
23º	22	2	5,498	**"bons"**
24º	5	2	5,483	
25º	4	2	5,462	
26º	12	2	5,180	
27º	38	2	4,934	
28º	17	2	4,931	
29º	19	2	4,502	
30º	15	4	4,637	
31º	14	4	4,511	
32º	33	4	4,157	
33º	37	4	3,912	
34º	45	4	3,898	
35º	6	4	3,497	
36º	42	4	3,422	**"regulares"**
37º	3	4	3,299	
38º	7	4	3,294	
39º	26	4	3,258	
40º	34	4	2,911	
41º	1	4	2,606	
42º	47	4	2,450	
43º	29	4	2,439	
44º	21	3	1,725	
45º	32	3	1,573	
46º	2	3	1,569	
47º	23	3	1,195	**"ruins"**
48º	16	3	0,494	
49º	48	3	0,318	
50º	44	3	0,003	

Assim, o Cluster 1**, fica melhor** com 16 vendedores que serão classificados como "ótimos", depois vem os demais na classificação:
Cluster 2, com 13 vendedores classificados como "bons";
Cluster 4, com 14 vendedores classificados como "regulares";
Cluster 3, com 7 vendedores classificados como "ruins".

A alocação dos vendedores em cada cluster se dá pelo cálculo da distância Euclidiana mínima em relação ao vetor peso dos neurônios.

O ranking final obtido pelo método híbrido (Tabela VII) e o ranking inicial obtido por Análise Fatorial (Tabela IV) apresentam coeficiente de correlação de Spearman igual a 0,99798 com valor-p = 0 que, pelos padrões normais, a associação entre as duas variáveis (resultados) seria considerada estatisticamente significante. Existe, portanto, consistência entre os métodos. Observa-se que a classificação obtida na formação dos clusters é praticamente a mesma da obtida na Análise Fatorial. Apenas as posições 29º, 30º e 31º não estão idênticas. Na Análise Fatorial, as posições 29º, 30º e 31º são ocupadas pelos vendedores de números 15, 14 e 19, respectivamente. E, no modelo híbrido proposto, são ocupadas pelos vendedores de números 19, 15 e 14. O que garante uma correlação elevada.

Aplicando K-means em substituição ao SOM, neste mesmo banco de dados, obteve-se resultado semelhante. Estabelecendo a formação de 4 clusters, o cluster 1 melhor classificado ("ótimos") tem 15 vendedores, seguido pelo cluster 2 classificado como "bons" com 14 vendedores, cluster 4 classificado como "regulares" também com 14 vendedores e o último cluster (cluster 3) classificado como "ruins" com 7 vendedores.

Para uma comparação dos resultados obtidos com AF/SOM (Análise Fatorial e SOM) e AF/K-MEANS (Análise Fatorial e K-Means) utilizados nesta análise, passou-se a uma avaliação das classificações encontradas.

Na Tabela VIII, estão as frequências de vendedores classificados pelas duas formas: AF/SOM e AF/K-MEANS.

TABELA VIII
CLASSIFICAÇÃO POR AF/SOM E AF/K-MEANS

AF/K-MEANS	AF/SOM				Total
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	
Cluster 1	15	0	0	0	15
Cluster 2	1	13	0	0	14
Cluster 3	0	0	7	0	7
Cluster 4	0	0	0	14	14
Total	16	13	7	14	50

Pode-se perceber que dos 50 vendedores classificados, 49 foram classificados da mesma forma. Os clusters 3 e 4 são idênticos aos gerados por AF/SOM. A distinção está nos clusters 1 e 2, em que a posição do vendedor número 24, que no AF/SOM está no cluster 1, no AF/K-MEANS está no cluster 2. O Coeficiente de Concordância de Kappa, estatisticamente significativo (p=0,000) foi de $\widehat{K} = 0{,}9727$, indicando uma concordância quase perfeita entre os agrupamentos utilizando SOM e K-MEANS.

O tempo de execução dos algoritmos foi calculado utilizando a função *Sys.time()* do Software R. Para AF/SOM o tempo de execução foi de 2.821008 segundos e para AF/K-MEANS foi de 1.096915 segundos.

VI. CONCLUSÃO

O estudo propôs um modelo híbrido de ranqueamento ou classificação, baseado nos resultados obtidos pela Análise Fatorial. A Análise Fatorial por meio da ponderação de seus escores fatoriais, permite a construção de um índice classificatório, em que a unidade com melhor desempenho na atividade fica melhor classificada. A Rede Neural SOM foi treinada utilizando como dado de entrada os valores dos escores fatoriais das variáveis latentes, estas novas variáveis preservam as informações dos dados originais. O propósito de se utilizar os escores das variáveis latentes é o de agrupar em clusters os indivíduos que são semelhantes nessa classificação.

Por meio do estudo de caso, conclui-se que o modelo híbrido desenvolvido apresenta resultados excelentes, visto que, após o treinamento da Rede Neural os resultados obtidos na classificação SOM, se comparado ao obtido da classificação inicial gerada pela Análise Fatorial, apresentaram coeficiente de correlação de Spearmann de 99%, sendo este um ótimo resultado. Para comparação foi aplicado o método K-MEANS em substituição ao SOM, os resultados são bem semelhantes, gerando coeficiente de concordância $\widehat{K} = 0{,}9727$, o que é quase perfeito.

REFERÊNCIAS

[1] R. Esquarcini and J. M. Marques. *"Classificação dos municípios paranaenses Segundo suas políticas setoriaispela análise multivariada."* Revista da FAE, vol .9, pp. 83-93, 2006.

[2] E. M. Furtado, A. Chaves Neto and Z. H. Domingues. *"Ranqueamento de faxinais do estado do Paraná através da Análise Fatorial*". Revista Ciências exatas e Naturais – RECEN. Unicentro, vol .5, no.1, 2003.

[3] Q. Ma, W. Wang, Q. Yao, J. Zhou e L. Quo. *"Factor analysis on call detail record."* In Proceedings of the 2018 27 th Wireless and Opyical Communication Conference (WOCC 2018), Hualien, Taiwan, pp. 1-5, 30 April-1 May 2018.

[4] M. Cottrell and M. Verleysen. *"Advances in Self-Organizing Maps."* Neural Networks 19, pp. 721-722, 2006.

[5] L. Churilov and A. Flitman. *"Towards fair ranking of Olympics achievements: the case of Sidney 2000."* Computers & Operational Research, vol.33, pp. 2057-2082, 2004.

[6] C. Spearman. *"General intelligence objectively determined and measured."* American Journal of Psychology, vol.15, pp. 201-293, 1904.

[7] L. L. Thurstone. "The vector of mind." Chicago: University of Chicago, 1935.

[8] A. S. Kaplunovsky. *"Why using factor analysis? dedicated to the centenary of factor analysis. 2009."* [Online]. Available: http://www.magniel.com/fa/kaplunovky.pdf. Accessed on: Nov. 5, 2019.

[9] F. H. C. Marriot. "The interpretation of multiple observations." New York, Academic Press, p. 117 1974.

[10] L. S. Kubrusly. "O Modelo de Análise Fatorial. Dissertação de Mestrado. UFRJ, 1981.

[11] J. F. Hair, R. E. Anderson, R. L. Tathan and W. C. Black. "Multivariate Data Analysis." 6 th ed. New Jersey: Prentice Hall, Ed. 2006.

[12] R. A. Johnson and D. W. Wichern. "Applied Multivariate Statistical Analysis." 6 th ed. New Jersey: Prentice Hall, Ed. 2007.

[13] J. Maroco. "Análise Estatística com utilização do SPSS." 3 rd ed. Lisboa: Edições Sílabo. P. 822 2007.

[14] L. V. Fauset. "Fundamentals of Neural Networks: Architectures, Algorithms, and Aplications". New Jersey: Prentice Hall, Ed. pp. 246-287, 1994.
[15] A. Ultch and C. Vetter. *"Self-Organizing Feature Maps versus Statistical Clustering Methods: A Benchmark."* Research Report no. 9, Dep. Of Mathematics, University of Marburg, 1994.
[16] T. Kohonen. *"Self-Organizing formation of topologically correct feature maps."* Biological Cybernetics, vol. 43, pp. 59-69, 1982.
[17] H. Simon. *"Neural Networks – A Comprehensive foundation."* 2 nd ed. Canadá: Prentice Hall, Ed., 1999.
[18] L. O. L. Oliveira. "Mapas Auto-Organizáveis de Kohonen aplicados ao mapeamento de ambientes de robótica móvel." Dissertação de mestrado, Universidade Federal de Santa Catarina, 2001.
[19] Y. Kou, H. Cui and L. Xu. "The Application of SOM and K-Means Algorithms in Public Security Performance Analysis and Forecasting." In: Khachidze V., Wang T., Siddiqui S., Liu V., Cappuccio S., Lim A. (eds) Contemporary Research on E-business Technology and Strategy. iCETS 2012. Communications in Computer and Information Science, vol 332. Springer, Berlin, Heidelberg. 2012. https://doi.org/10.1007/978-3-642-34447-3_7
[20] L. Zinger. A. Gobet and T. Pommier. "*Two decades of describing the unseen majority of aquatic microbial diversity."* Mol Ecol. Apr;21(8):1878-96. Doi:10.1111/j.1365-294X.2011.05362.x. 2012.

D. Steffen possui graduação Licenciatura em Matemática (2007) pela Faculdade Estadual de Filosofia, Ciências e Letras de União da Vitória, e Mestrado em Métodos Numéricos em Engenharia (2010) pela Universidade Federal do Paraná. Atualmente é professor da Universidade Comunitária da Região de Chapecó. Suas áreas de interesse incluem aprendizado de máquina, mineração de dados e redes neurais artificiais.

A. Chaves Neto possui graduação em Engenharia Civil pela Universidade Federal do Paraná (1974), Mestrado em Estatística pela Universidade Estadual de Campinas (1985) e Doutorado em Engenharia Elétrica (Sistemas Estocásticos e Estatística) pela Pontifícia Universidade Católica do Rio de Janeiro (1991). Professor Titular Aposentado do DEST/UFPR e atualmente é Professor Senior do PPGMNE da Universidade Federal do Paraná. Tem experiência em Métodos Estatísticos Multivariados, Previsão de Séries Temporais, Métodos de Estatística Aplicada em Mineração de Dados, Engenharia da Qualidade, Métodos Computacionalmente Intensivos, Reconhecimento de Padrões e Segurança de Barragens.