Composição de especialistas locais para classificação de dados

Omar J. S. Santos * Armando Z. Milioni *

* Instituto Tecnológico de Aeronáutica (ITA)
Divisão de Engenharia Mecânica-Aeronáutica-
São José dos Campos, SP – Brasil – CEP: 12228-900
{omarmai, milioni}@ita.br

Abstract

In this paper we present a Mixture of Local Experts Model (MLEM) for data classification. The discriminant tools applied are Fisher's Discriminant Analysis, Logistic Regression and a non-parametric model called Extended DEA-DA (Sueyoshi, 2004). Using real data, we compare the results obtained with the MLEM, which requires data clusterization and solution investigation on each cluster, against results obtained with a more orthodox approach, which is classification over the entire data set. The main conclusion is that even though it seems to be a promising technique, the additional effort in building a MLEM does not assure better results.

Resumo

Este artigo tem por objetivo apresentar um modelo de Composição de Especialistas Locais (CEL) como instrumento para classificação de dados. As técnicas discriminantes empregadas são a Análise Discriminante de Fisher, Regressão Logística e Modelos não paramétricos denominados "Extended DEA-DA" (Sueyoshi, 2004). Com base em uma massa de dados real, comparamos os resultados obtidos através da utilização do modelo CEL, que exige a clusterização da massa de dados e a busca da solução em cada cluster obtido, contra os resultados obtidos da maneira ortodoxa, que é a da busca de solução sobre a massa de dados global. A principal conclusão é a de que, embora seja uma técnica promissora, o esforço adicional na obtenção de um modelo CEL não assegura melhores resultados.

Keywords: Mixture of Local Expert Models; Discriminant Analysis; Clustering; Extended DEA-DA

Title: Mixture of Local Experts Model for Data Classification

1 Introdução

A classificação de dados tem se constituído num assunto de interesse permanente e de uso muito abrangente. Técnicas de análise discriminante fornecem subsídios para a classificação de dados em grupos distintos. Implementando essas técnicas em regiões específicas do espaço de dados de um problema qualquer e posteriormente compondo os resultados obtidos em cada região na tentativa de melhor classificar um novo entrante, chegamos a um modelo de Composição de Especialistas Locais (CEL) (ver fundamentos do assunto em Jacobs et alli, 1991; Lima et alli, 2002 e Melo et alli, 2004). Essa composição pode ou não resultar numa melhoria nas classificações desejadas e esse é o tema que será abordado no presente trabalho.

Este artigo tem por objetivo apresentar um modelo de Composição de Especialistas Locais (CEL) como instrumento para classificação de dados. Com base em uma massa de dados real, comparamos os resultados obtidos através da utilização da CEL com os resultados obtidos por modelos de análise discriminante aplicados sobre a massa de dados global, verificando a ocorrência ou não de melhoria no número de classificações corretas.

Este artigo está estruturado da seguinte maneira:

Na Seção 2 abordamos noções gerais de Análise Discriminante. Apresentamos uma breve descrição das técnicas discriminantes empregadas neste trabalho, que são a Análise Discriminante de Fisher, Regressão Logística e modelos do tipo Extended DEA-DA (Sueyoshi, 2004).

Na Seção 3 apresentamos os fundamentos da constituição de uma Composição de Especialistas Locais (CEL), sua estrutura e funções utilizadas como fatores de ponderação da classificação final.

Na Seção 4 fazemos um estudo de caso usando dados reais, explorando uma aplicação do modelo CEL sobre um conjunto de 95 empresas classificadas como solventes ou insolventes. Mostramos a clusterização feita, o resultado dos modelos discriminantes utilizados, a transformação dos valores dos melhores modelos locais em medidas de pertinência ao grupo das empresas solventes através de escalas de conversão, a construção do modelo CEL e sua comparação com o modelo discriminante que obteve os melhores resultados na massa de dados completa, ou global.

Na seção 5 comentamos as conclusões desse trabalho e indicamos sugestões para trabalhos futuros.

2 Análise Discriminante

A Análise Discriminante (DA, do inglês *Discriminant Analysis*) serve para classificar casos em valores categóricos de uma variável dependente freqüentemente dicotômica, ou seja, que pode assumir valores 0 ou 1, o que equivale a identificar esses casos como pertencentes ou não pertencentes a um determinado grupo.

Muitas áreas do conhecimento utilizam técnicas de DA para classificação em grupos, tais como medicina, biologia, economia, sensoriamento remoto, interpretação de imagens e outras.

Para que possamos classificar indivíduos (pessoas, plantas, coisas ou tudo o que for objeto

de estudo) torna-se necessário obter uma função discriminante. Calibrada a partir de uma massa de dados previamente classificada, essa função discriminante serve como modelo para que um entrante novo, i.e., indivíduo que não sabemos a que grupo pertence, seja classificado em um determinado grupo.

Para o desenvolvimento de nosso trabalho escolhemos três modelos de Análise Discriminante: (i) a função discriminante linear de Fisher (1936) (FLDF, do inglês *Fisher's linear discriminant function*), por tratar-se de um modelo clássico amplamente citado na literatura, servindo como referência para a avaliação de resultados; (ii) regressão logística, por ter sido o método empregado por Scarpel (2000), que levantou os dados do estudo de caso e (iii) o modelo Extended DEA-DA, modelo de programação mista proposto por Sueyioshi (2004), visando termos um modelo não-paramétrico inserido no contexto.

2.1 Análise Discriminante de FISHER

Consiste em separar duas ou mais classes de objetos e prever a pertinência de um novo objeto a uma das classes. Para melhor entendimento vamos considerar o caso de existência de apenas duas classes,G_1 e G_2. Os objetos ou atributos são separados ou classificados mediante medidas baseadas em p variáveis, isto é, são associados a vetores do tipo $X' = [X_1, X_2, X_3, ..., X_p]$.

Fisher tinha por objetivo transformar as observações multivariadas $X's$ (ditas variáveis independentes) em observações univariadas $Y's$(ditas variáveis dependentes), tal que os $Y's$ das classes G_1e G_2 fossem distanciados das médias das dados tanto quanto possível.

A idéia básica é a de criar uma combinação linear das variáveis independentes de tal forma a definir a variável dependente.

Segundo Lam et al (2003), a FLDF se esforça em prover uma função linear pela qual se associam valores a dois ou mais atributos independentes, os quais são combinados produzindo uma simples pontuação de classificação. Esta pontuação é comparada a um valor de corte que separa os dois grupos, permitindo então estabelecer a relação de pertinência do indivíduo a um dos grupos. Temos, portanto, uma equação linear do tipo $L = b_1x_1 + b_2x_2 + ... + b_nx_n + c$, onde os coeficientes b_i são calculados de forma a maximizar a razão entre a variância entre os grupos e a variância entre os indivíduos do grupo e c é uma constante semelhante ao intercepto de uma regressão linear. A seguir, indivíduos de uma amostra, oriundos de novas observações, são classificados nos grupos tendo por base os valores de seus atributos, calculados pela equação discriminante.

Se consideramos um problema de classificação com um critério determinado e uma amostra com n observações de dois grupos, G_1 e G_2, cujos valores do critério estabelecido são conhecidos, podemos formular a FLDF, a partir da fórmula:

$$\left(\overline{a^1} - \overline{a^2}\right)' S^{-1} a \tag{1}$$

onde, $\overline{a^1}$ e $\overline{a^2}$ são os vetores médios da amostra de, respectivamente, G_1 e G_2, S é a matriz de covariância da amostra e a é o vetor de valores de uma observação (ou caso). A regra de classificação baseada nas amostras se dá da seguinte maneira:

Classifica-se um novo entrante caracterizado por a em G_1 se

$$\left(\overline{a^1} - \overline{a^2}\right)' S^{-1} a \geq \frac{1}{2}\left(\overline{a^1} - \overline{a^2}\right)' S^{-1}\left(\overline{a^1} + \overline{a^2}\right) \tag{2}$$

onde, $(\overline{a_1} - \overline{a_2})\prime$ é o vetor da diferença entre os vetores médios transposto e S^{-1} é inversa da matriz de covariância.

Caso contrário, o novo entrante é classificado em G_2.

Dessa forma, o novo entrante pode ser classificado em um dos grupos devido a uma função discriminante oriunda dos dados de calibração.

2.2 Modelo de Regressão Linear Logística

Consideremos um vetor p-dimensional X, de variáveis independentes que se relacionam com uma variável dependente ou de resposta y**,** podendo esta assumir valores 0 ou 1. Sendo β_i e α os parâmetros e havendo n casos considerados, a probabilidade P_i, referente ao caso i, de que a variável dependente assuma o valor 1 pode ser representada por (ver Pindyck, 1998):

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\alpha + \beta X_i)}} \tag{3}$$

onde $Z_i = \alpha + \beta X_i$.

Essa expressão é conhecida como função logística acumulada. A probabilidade de que a variável y assuma o valor 0 é dada por:

$$1 - P_i = \frac{e^{-Z_i}}{1 + e^{-Z_i}} \tag{4}$$

Fazendo o logaritmo de $P_i/1 - P_i$ o modelo pode ser expresso como uma função linear das variáveis independentes ou preditoras:

$$\log \frac{P_i}{1 - P_i} = Z_i = \alpha + \beta X_i \tag{5}$$

Segundo Gujarati (2000):

a) Enquanto Z_i varia de $-\infty$ a $+\infty$, P_i varia entre 0 e 1;

b) P_i não se relaciona linearmente com Z_i, sendo portanto não-linear com as variáveis independentes X_i, daí a necessidade de se fazer o logaritmo de $P_i/1 - P_i$, tornando esse logaritmo uma relação linear com X_i;

c) Embora Z_i seja linear em X_i, as probabilidades propriamente ditas não o são, divergindo de um modelo de probabilidade linear (MPL) onde as probabilidades aumentam linearmente com X_i e apresentam o inconveniente de poderem extrapolar o intervalo [0,1].

d) Uma vez estimados os parâmetros do modelo, podemos calcular a probabilidade de y assumir o valor 1 ou 0, discriminando dois grupos, uma vez estabelecido um valor de corte.

O método da máxima verossimilhança é adequado à estimação dos parâmetros quando dispomos de observações individuais do pertencimento ou não a um determinado conjunto. Detalhes desse método para estimação dos parâmetros do modelo para o caso geral com mais de uma variável independente podem ser encontrados nos trabalhos de Scarpel (2000) e Scarpel e Milioni (2001 e 2002).

2.3 Modelos do tipo EXTENDED DEA-DA

Trata-se de um método não-paramétrico proposto por Sueyioshi (1999, 2001 e 2004) que atua como função discriminante se valendo de dois estágios de desenvolvimento. No primeiro, os elementos são classificados em um dos dois grupos ou numa área de intersecção, composta de elementos que não puderam ser facilmente classificados nesse primeiro estágio. No segundo estágio os elementos da área de intersecção são estudados visando classificá-los em um dos dois grupos. A técnica desenvolvida por Sueyioshi utiliza recursos da Análise de Envoltória de Dados (DEA, do inglês *Data Envelopment Analysis*) dentro de uma formulação de Análise Discriminante.

Para caracterizarmos a estrutura analítica do primeiro modelo DEA-DA de Sueyioshi (1999), vamos visualizar uma estrutura de DA e sintetizar o procedimento do modelo.

Como em DEA, sejam n DMU's j (do inglês, *Decision Making Units*; $j = 1, ..., n$) e observações com k fatores independentes i *(i=1,2,...,k)* que caracterizam seu desempenho denotado aqui por Z_{ij}. A análise discriminante pressupõe um conhecimento prévio de tal maneira que a partir de suas observações i, cada DMU j, possa ser classificada no grupo 1 (G_1) ou no grupo 2 (G_2). Tais grupos possuem, respectivamente, n_1 e n_2 observações. Como $G_1 \cap G_2 = \emptyset$ e $G_1 \cup G_2 = G$(conjunto de todas as DMU's), então $n_1 + n_2 = n$.

O primeiro modelo DEA-DA foi mais tarde alterado por Sueyoshi (2001) para que pudesse lidar com dados negativos, comuns em análises financeiras, sendo chamado a partir dessa alteração de Extended DEA-DA. Sueyoshi (2004) alterou novamente o modelo para que o segundo estágio do processamento minimizasse o número absoluto de classificações incorretas e ocorresse uma melhoria na eficiência computacional. É esse último modelo de Sueyoshi (2004) que empregamos neste trabalho.

O primeiro estágio desse modelo é formulado da seguinte maneira:

$$
\begin{array}{l}
\min s \\
\text{sujeito a:} \\
\sum_{i=1}^{k} (\lambda_i^+ - \lambda_i^-) Z_{ij} - d + s \geq 0, j \in G_1 \\
\sum_{i=1}^{k} (\lambda_i^+ - \lambda_i^-) Z_{ij} - d - s \leq 0, j \in G_2 \\
\sum_{i=1}^{k} (\lambda_i^+ + \lambda_i^-) = 1 \\
d, s : \text{irrestrito}; \zeta_i^+, \zeta_i^- : 0 \text{ ou } 1; \\
\lambda_i^+ \geq 0; \lambda_i^- \geq 0; \\
\text{NLC:(7),(8);NZC:(10)}
\end{array}
\tag{6}
$$

onde dé um valor limite, ou limiar, s representa um desvio e λ_i^+ e $\lambda_i^-, i = (1, 2...k)$ são pesos cujo papel passamos a explicar.

Foram definidas as seguintes variáveis:

$\lambda_i^+ = (|\lambda_i| + \lambda_i)/2$ e $\lambda_i^- = (|\lambda_i| - \lambda_i)/2$, para $i = 1, ..., k$

Trabalhando algebricamente as definições acima temos as seguintes conseqüências $|\lambda_i| = \lambda_i^+ + \lambda_i^-$ e $\lambda_i = \lambda_i^+ - \lambda_i^-$. Das definições, constatamos a condição de não linearidade ($\lambda_i^+ \lambda_i^- = 0$), uma vez que $\lambda_i^+ \lambda_i^- = (|\lambda_i|^2 - \lambda_i^2)/4 = 0$. Tal condição exclui a possibilidade de termos, simultaneamente, $\lambda_i^+ > 0$ e $\lambda_i^- > 0$.

A separação da variável λ_i em λ_i^+ e λ_i^- torna possível trabalhar não somente com dados positivos, mas também com dados negativos.

Especial atenção foi dada à condição de não linearidade (NLC, do inglês, *nonlinear condition*) e sua equivalência em programação mista (MIP, do inglês, *mixed integer programming*). Essa condição ($\lambda_i^+ \lambda_i^- = 0$) foi formulada introduzindo restrições com as variáveis bináriasζ_i^+ e ζ_i^-, da seguinte maneira:

$$\zeta_i^+ \geq \lambda_i^+ \geq \varepsilon \zeta_i^+ e \zeta_i^- \geq \lambda_i^- \geq \varepsilon \zeta_i^- \tag{7}$$

$$\zeta_i^+ + \zeta_i^- \leq 1, (i = 1, ..., k) \tag{8}$$

onde ε é um número muito pequeno, no estudo do autor foi utilizado $\varepsilon = 0,0005$.

As desigualdades em (7), que na formulação apresentada em (6) são referenciadas como NLC:(7), estabelecem os limites superior e inferior de λ_i^+ e λ_i^-. Em (8), referenciado em (6) como NLC:(8), temos que a soma das variáveis binárias é menor ou igual a um. Percebe-se que se tivéssemos $\lambda_i^+ \geq \varepsilon > 0$ e $\lambda_i^- \geq \varepsilon > 0$ em (7), então encontraríamos $\zeta_i^+ + \zeta_i^- = 2$ em (8), o que seria uma solução inviável. Portanto, $\lambda_i^+ > 0$ e $\lambda_i^- > 0$ não podem ocorrer simultaneamente.

Outra situação imposta é a condição de não nulidade (NZC, do inglês, *nonzero condition*), conforme estabelecida abaixo e que é referenciada em 6 como NZC:(9):

$$\sum_{i=1}^{k} (\zeta_i^+ + \zeta_i^-) = k \tag{9}$$

visando evitar $\lambda_i^+ = 0$ e $\lambda_i^- = 0$, simultaneamente. Tal condição impossibilita a desconsideração de uma variável ou fator significativo, o que ocorreria caso fosse possível $\lambda_i = \lambda_i^+ - \lambda_i^- = 0$.

Sejam $\lambda_i^*(= \lambda_i^{+*} - \lambda_i^{-*})$, d^* e s^* as soluções ótimas de (6). Se $s^* < 0$ não há área de intersecção entre os elementos dos dois conjuntos, i.e., todas as observações são claramente classificadas em G_1 e G_2. Se $s^* \geq 0$, existe uma área de intersecção e todos os dados são classificados num dos subconjuntos abaixo:

$$C_1 = \left\{ j \in G_1 / \sum_{i=1}^{k} \lambda_i^* z_{ij} > d^* + s^* \right\},$$
$$C_2 = \left\{ j \in G_2 / \sum_{i=1}^{k} \lambda_i^* z_{ij} < d^* - s^* \right\},$$
$$D_1 = G_1 - C_1,$$
$$D_2 = G_2 - C_2$$

A figura 1 mostra a separação nos quatro subconjuntos mencionados. Observamos que a área de intersecção corresponde a $D_1 \cup D_2$.

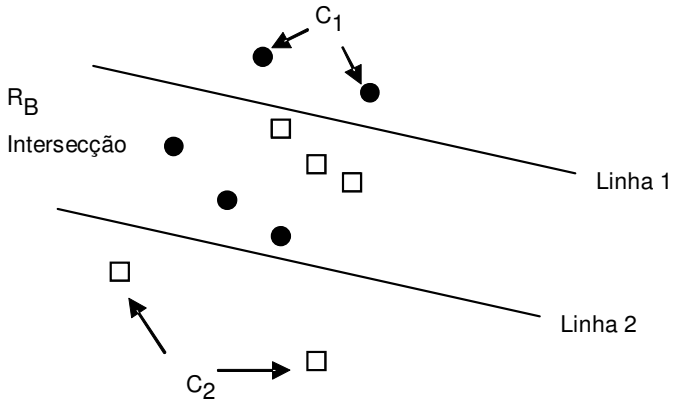

Figura 1: Classificação no primeiro estágio.

Matematicamente, três regiões são definidas no espaço como segue:

$$R_1 = \left\{ (z_1 ... z_n)^T / \sum_{i=1}^{k} \lambda_i^* z_i > d^* + s^* \right\},$$
$$R_2 = \left\{ (z_1 ... z_n)^T / \sum_{i=1}^{k} \lambda_i^* z_i < d^* - s^* \right\} e$$
$$R_B = \left\{ (z_1 ... z_n)^T / d^* - s^* \leq \sum_{i=1}^{k} \lambda_i^* z_i \leq d^* + s^* \right\}$$

Na figura 3, R_1 é o espaço de dados acima da linha 1 ($\lambda^* Z = d^* + s^*$). R_2, o espaço de dados abaixo da linha 2 ($\lambda^* Z = d^* - s^*$). A área de intersecção R_B se encontra entre as linhas 1 e 2.

No segundo estágio, para tratamento dos dados da área de intersecção, temos a formulação (10).

Nessa formulação Mé um número muito grande, como no conceito de Big – M em programação linear.

Neste modelo, a variável y_j indica a ocorrência de uma classificação incorreta e a função objetivo minimiza o número total de classificações incorretas.

$$\begin{aligned}
&\min \sum_{j \in D_1} y_j + \sum_{j \in D_2} y_j \\
&\text{sujeito a:} \\
&\sum_{i=1}^{k} (\lambda_i^+ - \lambda_i^-) z_{ij} - c + M y_j \geq 0, j \in D_1 \\
&\sum_{i=1}^{k} (\lambda_i^+ - \lambda_i^-) z_{ij} - c - M y_j \leq -\varepsilon, j \in D_2 \\
&\sum_{i=1}^{k} (\lambda_i^+ + \lambda_i^-) = 1 \\
&c : irrestrito; \zeta_i^+, \zeta_i^-, y_j = 0 \text{ ou } 1; \\
&\lambda_i^+ \geq 0; \lambda_i^- \geq 0 \\
&\text{NLC} : (7), (8); \text{NZC} : (9)
\end{aligned} \tag{10}$$

Nesse modelo, NLC (7) e (8) e NZC(9) repetem as equações da formulação apresentada em (6). Obtendo as soluções ótimas da formulação acima $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_k^*)$ e c^*, onde c é o

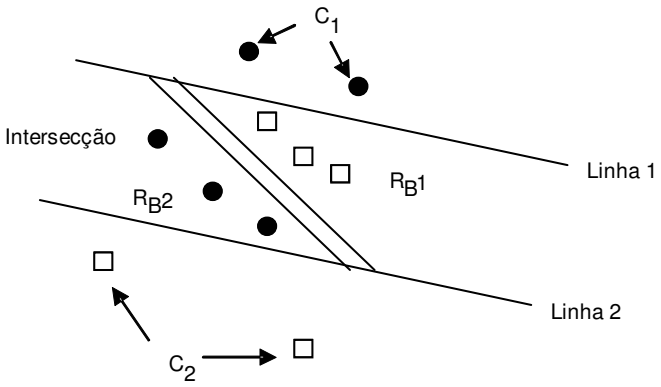

Figura 2: Classificação no segundo estágio.

valor discriminante no segundo estágio, a área de intersecção (R_B), identificada no primeiro estágio, pode ser separada da seguinte maneira:

$$R_{B1} = \left\{ (z_1 ... z_n)^T / \sum_{i=1}^{k} z_i \lambda_i^* \geq c^* \right\} \cap R_B,$$
$$R_{B2} = \left\{ (z_1 ... z_n)^T / \sum_{i=1}^{k} z_i \lambda_i^* \leq c^* - \varepsilon \right\} \cap R_B$$

A figura 2 mostra a classificação no segundo estágio.

Sintetizando, no primeiro estágio o modelo divide os dados em três grupos: G_1, G_2 e uma área ainda indefinida, chamada área de intersecção. No segundo estágio, os dados contidos na área de intersecção sofrem novo tratamento, sendo finalmente classificados em G_1 e G_2.

3 Composição de especialistas locais

A idéia básica de uma Composição de Especialistas Locais (CEL) para classificação de dados consiste em clusterizar uma massa de dados, aplicar diferentes técnicas discriminantes ditas "modelos especialistas" em cada clusters, ponderar os resultados das técnicas discriminantes vencedoras, que são aquelas com o maior número de classificações corretas em cada cluster, e obter um valor numérico que permita classificar uma observação nova (novo entrante) como pertencente ou não a um determinado grupo.

Aqui cabe levantar uma questão importante. Cada modelo utilizado em análise discriminante gera resultados numéricos que, segundo um critério estabelecido, permite classificar as observações em grupos. A natureza do valor numérico gerado, todavia, difere de modelo para modelo e até mesmo dentro de um mesmo modelo, como é o caso dos modelos de dois estágios de Sueyoshi, em que o valor numérico obtido na análise do segundo estágio não guarda relação com aquele obtido no primeiro estágio. Para contornar a dificuldade de composição desses valores de natureza distinta, converteremos os valores numéricos gerados em medidas que representam o grau de pertinência de uma determinada observação a um determinado grupo. Essa conversão será detalhada na seção 4.3, adiante.

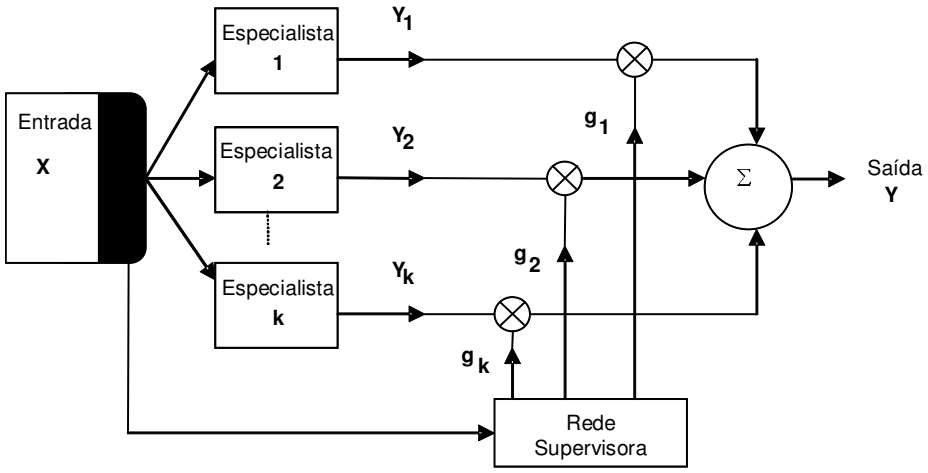

Figura 3: Composição de Especialistas locais.

A figura 3 ilustra o conceito de CEL. Nela, uma massa de dados X foi dividida em k clusters. Em cada cluster houve uma técnica discriminante com melhor desempenho (modelo especialista vencedor). Cada modelo vencedor gera um uma saída Y_i que é transformada numa medida de grau de pertinência a um grupo. As diversas saídas Y_i são ponderadas por uma função gerando uma saída única Y que define a classificação final.

A saída Y é dada por:

$$Y = \sum_{i=1}^{k} g_i y_i \tag{11}$$

Para cálculo do fator de ponderação g_i utilizamos o mesmo procedimento de Melo (2003), que se baseia na distância d_i, definida a seguir:

$$d_i = \exp\left[-\frac{1}{2(s_i^2/S^2)}\left\|x - ctr_i\right\|^2\right] \tag{12}$$

onde:

s_i^2 é a variância do cluster i,

S^2 é a maior variância apresentada pelos clusters, isto é, $S^2 = \underset{i}{Max}(s_i^2)$ e

$\|x - ctr_i\|$ é a distância euclidiana da entrada x ao centro do cluster i.

Uma vez calculado o valor de d_i, definimos g_i do seguinte modo:

$$g_i = \frac{d_i}{\sum\limits_{i=1}^{M} d_i} \tag{13}$$

Dessa forma para M clusters temos que $\sum\limits_{i=1}^{M} g_i = 1$.

Tabela 1: Centróides de três clusters

	Cluster No. 1	Cluster No. 2	Cluster No. 3
GA	2,2930	6,1659	0,7640
RA	0,1415	0,7059	-0,1778

Tabela 2: Composição dos clusters obtidos

Empresa	Cluster 1	Cluster 2	Cluster 3	Total
insolv	2	0	31	33
solv	33	9	20	62
Total	35	9	51	95

4 Estudo do caso

Em nosso estudo de caso investigamos a calibração de um modelo de composição de especialistas locais (CEL) para classificar empresas em dois conjuntos: G_1 (insolventes) e G_2(solventes). A massa de dados utilizada é a mesma de Scarpel (2000) e Almeida (2000). Ela é composta por 95 empresas, dentre as quais 33 são insolventes e 62 são solventes. Todas são empresas de capital aberto cujas demonstrações financeiras estavam disponíveis na Comissão de Valores Mobiliários (CVM) e na BOVESPA (Bolsa de Valores de São Paulo). Como variáveis explicativas, ficaremos com a mesma escolha de Almeida (2000), que foi a seguinte:

GA – Índice de Giro do Ativo Total, resultado da relação entre receita anual (vendas) e ativo total, dividido pelo Índice de Endividamento Geral, resultado da relação entre o exigível total (= passivo circulante + exigível a longo prazo) e o ativo total;

RA – Taxa de Retorno sobre o Ativo Total, resultado da relação entre o lucro (antes do pagamento de juros + imposto de renda) e o ativo total, dividido pelo Índice de Endividamento Geral.

Para a clusterização, estimação da FLDF e da regressão logística, empregamos o software Statistica, versão 5.5 (1999).

4.1 Clusterização

As 95 (noventa e cinco) empresas, foram clusterizadas de maneira a agrupá-las por similaridade. Após um estudo de diversas alternativas quanto ao número k de clusters (ver Santos, 2004), optamos por trabalhar com 3 clusters. Na figura 4 podemos visualizar os clusters obtidos.

A tabela 1 apresenta os centróides dos 3 clusters obtidos.

A tabela 2 resume a composição, i.e., o número de empresas solventes e insolventes em cada um dos 3 clusters obtidos.

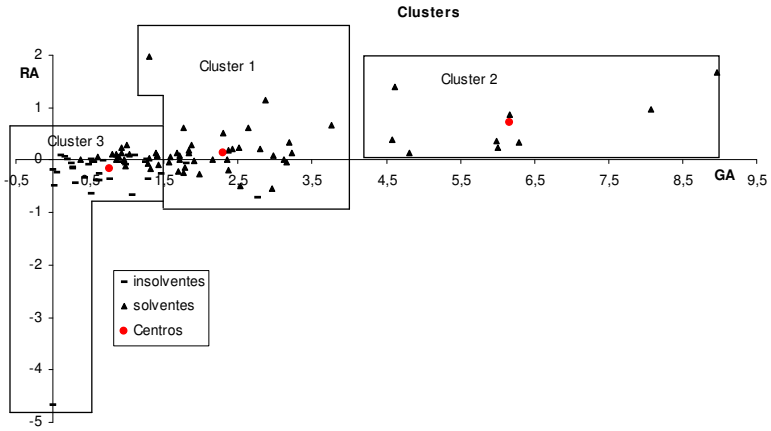

Figura 4: Clusterização em três grupos

4.2 Resultados dos Modelos Discriminantes

Aplicamos a Análise Discriminante de Fisher, Regressão Logística e o modelo Extended DEA-DA na massa de dados global (i.e., sem clusterização) para verificar qual modelo discriminante apresentaria o maior número de classificações corretas. Esse é o nosso modelo vencedor global e constitui o modelo de referência para comparação com os resultados da utilização do modelo CEL. A tabela 3 resume os resultados obtidos.

Tabela 3: Resultado na massa de dados global

	Global			
acertos	insolv	solv	total	%
AD	31	43	74	77,9
Logit	28	58	86	90,5
DEA-DA	28	59	87	91,6

Tabela 4: Resultados obtidos no cluster 3

acertos	insolv	solv	total	%
AD	22	18	40	78,4
Logit	27	16	43	84,3
DEA-DA	26	18	44	86,3

Como podemos verificar, o modelo Extended DEA-DA foi o vencedor na massa de dados global e, portanto, é a referência de comparação com os resultados do modelo CEL.

No Cluster 1, que contém somente duas empresas insolventes, não é razoável aplicar qualquer modelo estatístico. Nesse contexto, descartamos a análise discriminante de Fisher e a regressão logística. Fizemos uma tentativa então com o modelo não-paramétrico Extended DEA-DA que, conseqüentemente, por ser o único, foi o modelo vencedor nesse cluster.

A calibração apresentou apenas uma empresa que, no segundo estágio, teve seu valor de discriminação situado entre os valores de referência $d^* + s^*$ e $d^* - s^*$. Na impossibilidade de definir a pertinência a um dos dois grupos, consideramos essa classificação como errada. Portanto, o modelo apresentou apenas um erro de classificação e um percentual de acerto de 97,1%.

O Cluster 2 apresenta somente nove empresas solventes, não sendo necessário qualquer esforço de discriminação. À qualquer empresa desse cluster atribuímos 100% de pertinência a G_2(solventes).

O Cluster 3 nos permite trabalhar com todos os modelos especialistas considerados.

A tabela 4 resume os resultados obtidos pelos modelos especialistas aplicados ao Cluster 3, o qual contém 31 empresas insolventes e 20 solventes.

Assim, o modelo especialista vencedor para o cluster considerado foi o Extended DEA-DA.

Com isso, nossa composição se reporta a um único modelo aplicado a clusters diferentes, produzindo superfícies de separação e funções discriminantes distintas.

4.3 Escala de Conversão

Já vimos que o modelo CEL será composto por um único tipo de especialista local, o Extended DEA-DA. Um questionamento que aflora nesse ponto é o de como combinar os valores atribuídos a cada caso (empresa), uma vez que os mesmos apresentam ordem de grandeza distinta conforme tenham sido obtidos no primeiro ou no segundo estágio de classificação.

A dificuldade maior, quando da conversão dos valores atribuídos pelo modelo Extended

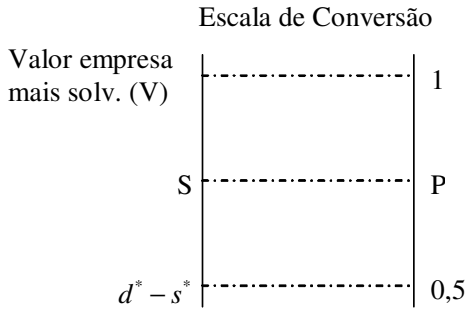

Figura 5: Escala de conversão em Pertinência (Extended DEA-DA)

DEA-DA, recai no fato de termos dois estágios e, portanto, duas escalas distintas. Não há qualquer conexão entre os valores atribuídos no primeiro estágio e os valores do segundo estágio. No entanto, tais escalas não devem apresentar comportamentos independentes, ou poderíamos ter casos em que uma empresa que não pode ser classificada em um dos grupos no primeiro estágio, por ter se localizado na área de intersecção, registraria um grau de pertinência maior do que uma empresa que foi classificada no primeiro estágio. Isso equivaleria a dizer que a segurança na classificação da empresa que apresentou dúvida no primeiro estágio é maior do que o daquela para a qual não houve dúvida, o que não parece ser lógico.

Um cuidado essencial nessa conversão é o fato de que os valores percentuais obtidos, quando comparados a um determinado limiar, devem refletir exatamente as classificações obtidas pelo modelo especialista antes da conversão.

Nesse contexto as escalas devem apresentar coerência e representar fielmente a classificação atingida pelo modelo. Para contornar tais problemas adotamos a seguinte solução. Para o primeiro estágio, o valor inferior da área de intersecção $(d^* - s^*)$, que contém os pontos que terão sua classificação definida apenas no segundo estágio, foi arbitrado um valor de pertinência P ao grupo das empresas solventes igual a 0,5. Ao maior valor atribuído pelo modelo, que corresponde à empresa, digamos assim, mais claramente solvente, foi arbitrado o valor 1. Montamos então a escala de conversão ilustrada pela figura 5 e expressa pela relação dada em (14):

$$\frac{\mathrm{S} - (d^* - s^*)}{\mathrm{V} - (d^* - s^*)} = \frac{\mathrm{P} - 0,5}{1 - 0,5} \tag{14}$$

onde S é o valor atribuído pelo modelo Extended DEA-DA à empresa em questão, V é o valor atribuído pelo modelo à empresa "mais claramente solvente" e P é o valor de pertinência a ser obtido para a empresa em questão.

Todavia, quando aplicada a novos entrantes, essa escala poderá apresentar distorções, já que, por basear-se em uma amostra, não há garantias de que o valor de P esteja entre 0 e 1. Para que tais valores possam ser vistos como a probabilidade de pertencer a um grupo, utilizamos a solução proposta por Gujarati (2000), limitando em zero os valores de pertinência inferiores a zero e em um os valores de pertinência superiores a um. Dessa forma, esses valores de pertinência podem ser vistos como probabilidades.

A expressão de conversão para o segundo estágio é a seguinte:

$$\mathrm{P} = (P_{ref} - 0,5)\left(\frac{\mathrm{S} - c^*}{\theta.(Ic^*)}\right) + 0,5 \quad (15)$$

onde P_{ref} é a probabilidade do caso de referência (classificado como solvente) no primeiro e segundo estágios (com valor mais próximo de c^*), S é o valor atribuído no segundo estágio do modelo Extended DEA-DA, c^* é o limiar do segundo estágio, θ é um parâmetro que visa a adequação da escala e I é uma função indicadora que poderá assumir os valores 1 e -1. Essa função indicadora será utilizada somente para adequação do sinal, lembrando que uma empresa para ser considerada solvente deve apresentar valor maior do que 0,5. No caso prático estudado arbitramos $\theta = 0,05$.

4.4 Modelo CEL

O cálculo das ponderações do modelo CEL se dá de acordo com as expressões (13), (14) e (15).

A saída y_i é a probabilidade de pertinência ao grupo das empresas solventes(G_2), resultado da conversão em probabilidades dos valores atribuídos em cada cluster.

Vamos ilustrar o cálculo completo para a empresa de número 95, escolhida ao acaso, que é solvente, pertence ao Cluster 1 e para a qual GA = 1,913 e RA = -0,009.

Calculando a variância de cada cluster, obtemos os seguintes valores para os Clusters 1, 2 e 3:

$$s_1^2 = 0,1523, \quad s_2^2 = 0,8768 \quad e \quad s_3^2 = 0,3661.$$

Como a maior variância é a do cluster 2, temos $S^2 = 0,8768$.

Temos ainda que:

$$\|x_{95} - ctr_1\|^2 = 0,167, \quad \|x_{95} - ctr_2\|^2 = 18,598 \quad e \quad \|x_{95} - ctr_3\|^2 = 1,349,$$

assim, encontramos:

$$d_1 = 0,6183, \quad d_2 = 0,0001 \quad e \quad d_3 = 0,1989,$$

o que nos leva a:

$$g_1 = 0,7565, \quad g_2 = 0,0001 \quad e \quad g_3 = 0,2434.$$

Os modelos locais vencedores em cada cluster aplicados aos dados da empresa 95 geram saídas que, convertidas pela escala apresentada em 4.3, transformam-se nas seguintes probabilidades de pertinência ao grupo das empresas solventes:

$$\mathrm{P}_{c1} = 0,5313, \quad \mathrm{P}_{c2} = 1 \quad e \quad \mathrm{P}_{c3} = 0,6714.$$

Então, calculamos a seguinte probabilidade para o modelo CEL:

$$\mathrm{P}_{CEL} = g_1\mathrm{P}_{c1} + g_2\mathrm{P}_{c2} + g_3\mathrm{P}_{c3} = 0,5654$$

Como esse número é superior a 0,5 a empresa 95 é classificada como solvente.

Tabela 5: Comparação entre modelo CEL e Extended DEA-DA

acertos	insolv	solv	total	%
CEL	29	58	87	91,6
DEA-DA	28	59	87	91,6

Uma vez calculados os valores para todas as empresas, resta-nos comparar os resultados do modelo CEL com o resultado do especialista vencedor na massa de dados global. A tabela 5 resume a comparação de resultados.

Verificamos que, para a massa de dados estudada, não houve melhoria no número de classificações corretas ao adotarmos o modelo CEL, se comparado ao modelo Extended DEA-DA aplicado sobre a massa de dados global. Ambos registram um percentual de acerto de aproximadamente 91,6%. O modelo Extended DEA-DA registra 5 empresas insolventes e 3 empresas solventes incorretamente classificadas. Já o modelo CEL registra 4 empresas insolventes e 4 empresas solventes incorretamente classificadas.

5 Conclusões

Neste trabalho abordamos aspectos relativos a técnicas de análise discriminante e construção de uma Composição de Especialistas Locais (CEL) para classificação de dados. Para isso, fizemos uso de três técnicas de discriminação, a saber, Análise Discriminante de Fisher, Regressão Logística e Extended DEA-DA.

No decorrer do desenvolvimento, definimos o caso estudado, no qual apresentamos uma massa de dados onde 95 empresas se enquadravam na categoria solvente ou insolvente. Essa massa de dados foi clusterizada e tornou-se a base da calibração do nosso modelo CEL. Os resultados obtidos indicaram o modelo Extended DEA-DA como único vencedor, tanto na massa de dados global quanto na massa de dados clusterizada, exceto no cluster constituído somente de empresas solventes.

Um aspecto importante foi a necessidade da construção da escala de conversão de valores do modelo discriminante para graus de pertinência ao grupo de empresas solventes. Nesse aspecto não vislumbramos uma solução geral, acreditamos tratar-se de um problema prático que deverá ser contornado caso a caso, como fizemos no nosso estudo de caso.

Ao compararmos o modelo CEL com a técnica discriminante vencedora na massa de dados global, os números finais mostraram que ambos apresentaram idêntico número absoluto de classificações corretas, perfazendo um percentual aproximado de 91,6% de acerto na calibração.

Esse resultado indica que o esforço adicional empregado na partição da massa de dados em regiões e aplicação de soluções nessas regiões, que implica grande esforço adicional em comparação ao procedimento ortodoxo de aplicar a solução sobre a massa de dados global, não necessariamente assegura melhores resultados.

Como sugestões para trabalhos futuros podemos indicar:

- um estudo mais geral sobre a construção de escalas de conversão de valores dos modelos discriminantes em valores percentuais que representem graus de pertinência a um determinado

conjunto;

- estudar a adequação do uso do parâmetro subjetivo θ na conversão dos valores obtidos pelo modelo Extended DEA-DA em valores percentuais para outras massa de dados, utilizando simulação.

- a utilização de outras ferramentas de classificação de dados, redes neurais e outros especialistas, para obtenção de modelos CEL diferenciados.

- a aplicação de modelos CEL numa massa de dados maior, possibilitando separar parte dos dados para calibração e outra parte para teste, verificando-se assim a capacidade de generalização do modelo.

6 Referências

ALMEIDA, H. R. **Análise de envoltória de dados na tomada de decisão para concessão de crédito.** Dissertação (Mestrado em Produção) – Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, 2000

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, p.179-188, 1936

GUJARATI, D. N. **Econometria básica**. São Paulo: Makron Books, 2000

JACOBS, R. A.; JORDAN, M. I.; NOWLAN, S. J. & HINTON, G. E. Adaptive Mixture of Local Experts. **Neural Computation**. Vol. 3, No. 1, pp.79-87, MIT Press, 1991

LAM, K.F.; MOY, J.W. A piecewise linear programming approach to the two- group discriminant problem: an adaptation to Fisher's linear discriminant function model. **European Journal of Operational Research**, v.145, p. 471-481, 2003

LIMA, C. A. M.; COELHO, A. L. V.; VON ZUBEN, F. Mixture of Experts Applied to Nonlinear Dynamic Systems Identification:A Comparative Study, **Proceedings of the VII Brazilian Sympsium on Neural Networks**, Porto de Galinhas, Recife, Brazil, Nov 11-14, 2002, pp 162-167, 2002

MELO, B. **Previsão de séries temporais usando modelos de composição de especialistas locais.** Dissertação (Mestrado em Produção) - Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, 2003

MELO, B.; NASCIMENTO Jr, C. L.; MILIONI, A. Z.. **Daily Sugar Price Forecasting Using Mixture of Local Experts Models**. In: ZANASI, A.; EBECKEN, N.f.f.; BREBBIA, C.a. (Org.). **Data Mining V**: **Data Mining, Text Mining and their Business Applications**. Londres, v. 10, p.271-281, 2004

PINDYCK, R. S.; RUBINFELD, D. L. **Econometric models and economic forecasts**. 4. ed. New York: McGraw-Hill, 1998.

SANTOS, O. J. S. . **Composição de Especialistas Locais para Classificação de Populações.** Dissertação (Mestrado em Produção) - Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, 2004

SCARPEL, R. A. **Modelos matemáticos em análise financeira de empresas de setores industriais e de crédito**. Dissertação (Mestrado em Produção) – Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, 2000

SCARPEL, R. A.; MILIONI, A. Z.. Aplicação de modelagem econométrica à análise financeira de empresas. **Revista de Administração (RAUSP)**, São Paulo, SP, v. 36, n. 2, p. 80-88, 2001

SCARPEL, R. A.; MILIONI, A. Z.. Utilização conjunta de modelagem econométrica e otimização em decisões de concessão de crédito. **Pesquisa Operacional**, v. 22, n. 1, p.61-72, 2002

STATSOFT INK. STATISTICA 5.5, **Software Manual**, Tulsa, 1999

SUEYOSHI, T. DEA: discriminant analysis in the view of goal programming. **European Journal of Operational Research**, v.115, p. 564-582, 1999

SUEYOSHI, T. Extended DEA-discriminant analysis. **European Journal of Operational Research**, v.131, p. 324-351, 2001

SUEYOSHI, T. Mixed integer programming approach of extend DEA- discriminant analysis. **European Journal of Operational Research**, v.152, p.45-55, 2004