versão impressa ISSN 0101-7438 / versão online ISSN 1678-5142

OTIMIZAÇÃO NA FORMAÇÃO DE AGRUPAMENTOS EM PROBLEMAS DE COMPOSIÇÃO DE ESPECIALISTAS

Rodrigo Arnaldo Scarpel *
Engenharia Aeronáutica e Mecânica (EAM)
Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP
rodrigo@ita.br

Armando Zeferino Milioni
Engenharia Aeronáutica e Mecânica (EAM)
Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP
milioni@ita.br

* *Corresponding author* / autor para quem as correspondências devem ser encaminhadas

Recebido em 02/2006; aceito em 08/2006
Received February 2006; accepted August 2006

Resumo

A estimação de funções a partir de um conjunto limitado de amostras é um problema central em estatística aplicada. Um grande número de abordagens para tratar esse problema foi proposto como os métodos dos mínimos quadrados por Gauss e de mínimo módulo por Laplace, e, mais recentemente, o uso de redes neurais, de *support vector machines*, de composição de especialistas, dentre outros. Neste trabalho abordou-se a composição de especialistas e otimização na formação de agrupamentos, que engloba análise exploratória, mineração de dados e modelagem em uma única técnica, útil, por exemplo, na criação de modelos preditivos. A idéia básica da composição de especialistas é particionar o espaço de entrada em diferentes regiões e em cada região seleciona-se o especialista mais adequado. Propôs-se, então, a otimização na formação dos agrupamentos como uma forma de melhorar a qualidade dos ajustes dos modelos e das previsões realizadas.

Palavras-chave: mistura de especialistas; formação de agrupamentos; otimização.

Abstract

Estimation of real-valued functions from a finite set of samples is a central problem in applied statistics. Many different approaches to deal with this problem were proposed as the least-squares method by Gauss, the least-modulus method by Laplace, and more recently the usage of neural networks, support vector machines, mixture of local expert models, amongst others. We addressed the issues mixture of local expert models (MLEM) and clustering optimization, which congregates exploratory analysis, data mining and mathematical modeling in the same technique, used, for example, in the development of predictive models. The basic idea of MLEM is clustering the points from the entry data set, and then different modeling techniques are applied in order to select the best model for each cluster. We proposed a clustering optimization procedure as a way to improve the performance on both the fitting of the models and their usage in forecasting.

Keywords: mixture of experts; clustering; optimization.

1. Introdução

A estimação de funções a partir de um conjunto limitado de amostras é um problema central em estatística aplicada. Um grande número de abordagens para tratar esse problema foi proposto como os métodos dos mínimos quadrados por Gauss e de mínimo módulo por Laplace, no século XIX e, mais recentemente, o uso de redes neurais, de *support vector machines*, de composição de especialistas, dentre outros.

Em geral, sabe-se que nenhum dos métodos propostos é completo. Alguns apresentam alta velocidade de convergência, mas podem apresentar falhas de generalização detectadas na validação cruzada. Outros conseguem boa generalização ao custo de baixa velocidade de convergência.

A composição de especialistas é uma abordagem que tem por princípio combinar os resultados obtidos por vários tipos de especialistas, i. e., várias técnicas matemáticas usadas para a estimação de funções, sendo sua meta conseguir um resultado melhor do que aquele que seria obtido pelos especialistas trabalhando de forma isolada. As etapas dessa abordagem são: (a) dividir o espaço de atributos utilizando algum método de geração de agrupamentos, (b) designar um especialista para responder em cada região do espaço de atributos, (c) implementar a composição de especialistas usando uma rede supervisora que decide como ponderar as saídas de cada especialista.

O objetivo deste trabalho é propor a otimização na formação de agrupamentos, que consiste em integrar as etapas de formação de agrupamentos e de designação dos especialistas, em composição de especialistas aplicados a problemas de regressão, para melhorar a qualidade dos ajustes dos modelos e das previsões realizadas.

2. Composição de especialistas

A composição de especialistas (do inglês *Mixture of Local Expert Models*) é uma abordagem proposta por Jacobs *et al.* (1991) que consiste em usar as diferentes técnicas existentes com a estratégia "dividir para conquistar", decompondo problemas complexos em tarefas mais simples.

Segundo Melo *et al.* (2003) a característica central da composição de especialistas é o uso de vários tipos de especialistas, em que cada tipo de especialista usa uma técnica diferente de modelagem, assumindo que diferentes especialistas apresentam diferentes desempenhos em diferentes regiões do espaço de entrada.

Pinto (2003) indica que a composição de especialistas locais permite misturar modelos simples transformando-os em poderosas ferramentas para tratar problemas complexos e que os especialistas podem ser, por exemplo, modelos de regressão, que combinados através de um conjunto de pesos, podem auxiliar na obtenção de melhores soluções do que aquelas obtidas pelo modelo global, que é o modelo obtido a partir de todos os dados do conjunto de treinamento.

As etapas de construção de um modelo de previsão usando a composição de especialistas locais são:

1. Inicialmente o espaço de entrada (X) é particionado em várias regiões (X_c), utilizando um método de geração de agrupamentos;

2. Em cada região os diferentes modelos são treinados utilizando apenas os dados daquela região;
3. Para cada região é eleito o melhor especialista, segundo algum critério específico e esse especialista passará a ser considerado o único especialista daquela região.

2.1 Partição do espaço de entrada

Para executar a etapa inicial de partição do espaço de entrada qualquer um dos métodos de geração de agrupamentos pode ser utilizado, uma vez que todos os métodos existentes se propõem a descobrir similaridades nos dados colocados na sua entrada durante o treinamento. A partir deste princípio são obtidos agrupamentos (*clusters*) que estão próximos, segundo algum critério de distância, definindo uma região do espaço de entrada.

Dentre os métodos de geração de agrupamentos, destaca-se o algoritmo k-médias que é um método partitivo que tem como critério de erro a distância Euclideana. Se o objetivo é particionar n observações, no espaço p-dimensional, em k agrupamentos, a formulação, por programação matemática, do algoritmo k-médias é:

$$Min \sum_{i=1}^{n}\sum_{c=1}^{k} z_{ic} \left[\sum_{j=1}^{p}(x_{ij}-m_{cj})^2\right]^{\frac{1}{2}} \qquad (1)$$

$$Sujeito\ a \quad \sum_{c=1}^{k} z_{ic} = 1 \qquad i = 1,...,n$$

$$z_{ic} = \begin{cases} 1, \text{ se o ponto i pertencer ao agrupamento c} \\ 0, \text{ caso contrário} \end{cases} \qquad , c = 1,...,k \text{ e } i = 1,...,n$$

em que

$$m_{cj} = \frac{\sum_{i=1}^{n} z_{ic}\ x_{ij}}{\sum_{i=1}^{n} z_{ic}} \qquad c = 1,...,k \quad e \quad j = 1,...,p \qquad (2)$$

sendo m_{cj} o centróide do agrupamento c na dimensão j.

Desta forma, objetiva-se determinar em qual dos k agrupamentos cada um dos n pontos será alocado. A variável de decisão (z_{ic}) é binária recebendo valor 1 se o ponto i pertencer ao agrupamento c ou 0 caso contrário e cada ponto só pode pertencer a um agrupamento. A função objetivo do problema minimiza a soma da distância Euclideana entre os pontos e os centróides, no espaço R^p, e quando o ponto é alocado ao agrupamento c seus atributos passam a interferir no centróide desse agrupamento (m_{cj}).

Segundo Mangasarian (1997), esse problema pode ser resolvido por um procedimento de otimização iterativa em dois passos. No primeiro passo é feita a atribuição dos pontos aos agrupamentos e no segundo passo os centróides dos agrupamentos são atualizados, levando-se em consideração as alocações correntes. Isso conduz ao seguinte algoritmo:

Algoritmo k-médias – Dado os k centróides dos agrupamentos $m_{1,t}$, $m_{2,t}$, …, $m_{k,t}$ na iteração t, calcula-se $m_{1,t+1}$, $m_{2,t+1}$, …, $m_{k,t+1}$ pelos seguintes passos:

1. Alocação dos pontos aos agrupamentos: Para cada i = 1,…,n ponto, aloca-se x_i ao agrupamento c de forma que $m_{c(i),t}$ seja o centróide mais próximo de x_i tomando como critério a distância Euclideana.
2. Atualização dos centróides: Para c = 1,…,k faz-se $m_{c,t+1}$ ser a média de todos os x_i alocados a $m_{c,t}$. Para-se quando $m_{c,t+1} = m_{c,t}$, c = 1,…,k.

Em relação ao número de iterações necessárias para a convergência, Duda *et al.* (2001) afirmam que é muito menor do que o número de pontos existentes. Esses autores posicionam o algoritmo k-médias em uma categoria de procedimentos iterativos de otimização, pois os valores dos centróides tendem a se mover de forma a minimizar uma função de erro quadrática, podendo, então, ser vista como uma forma de se obter estimativas de máxima verossimilhança da média.

Segundo Webb (2002), apesar da solução obtida por esse procedimento iterativo ser sub ótima, o valor da função objetivo da solução obtida por esse método não é muito maior que o valor da função objetivo da solução ótima, sendo, portanto, uma boa aproximação da mesma. O autor afirma, ainda, que a necessidade computacional para se obter a solução ótima é proibitiva, mesmo nos casos com moderado número de pontos.

2.2 Treinamento dos modelos

Para a escolha dos especialistas locais os dados são divididos em dois conjuntos, denominados conjunto de treinamento e conjunto de validação. Desta forma, será eleito o especialista local aquele que obtiver as melhores estatísticas de desempenho, calculadas a partir do conjunto de validação.

Segundo Melo *et al.* (2003), em relação aos modelos treinados em cada região do espaço de entrada, é importante que sejam escolhidos de forma a representar diferentes técnicas de modelagem, como as que dão ênfase à não-linearidade como, por exemplo, a análise de regressão não-linear e à linearidade como, por exemplo, a análise de regressão linear.

A etapa de treinamento dos modelos pode ser vista como um problema de otimização em que o objetivo é buscar, para cada tipo de modelo, os parâmetros que minimizem o risco empírico (Vapnik, 1999) dado por

$$R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) \tag{3}$$

em que $L(Y_i,f(X_i))$ é uma função de perda. Assim, a formulação do problema de otimização para a estimação dos parâmetros do modelo é

$$Min \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) \tag{4}$$

sendo que, no caso linear, $f(X_i) = \alpha + \beta.x_i$ e, no caso quadrático, $f(X_i) = \beta_0 + \beta_1.x_i + \beta_2.x_i^2$, em que α e β são os parâmetros a serem estimados no caso linear e β_0, β_1 e β_2 são os parâmetros a serem estimados no caso quadrático.

As funções de perda mais comumente utilizadas em problemas de regressão são a de mínimos quadrados, dada por

$$L(y, f(x)) = (y - f(x))^2 \tag{5}$$

a de mínimo módulo, dada por

$$L(y, f(x)) = |y - f(x)| \tag{6}$$

e a robusta (Huber, 1964), dada por

$$L(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & \text{se } |y - f(x)| < \mu \\ \mu|y - f(x)| - \frac{\mu^2}{2}, & \text{caso contrário} \end{cases} \tag{7}$$

em que μ é um parâmetro arbitrado.

Vapnik (1995) propôs uma nova função de perda chamada ε-insensitiva dada por

$$L_\varepsilon(y, f(x)) = \begin{cases} 0, & \text{se } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon, & \text{caso contrário} \end{cases} \tag{8}$$

Essa função de perda foi proposta no contexto da regressão *support vector machine* (SVM). O princípio desta função de perda é penalizar apenas o erro maior que um valor arbitrado (ε).

3. Otimização na formação dos agrupamentos

Pela abordagem de composição de especialistas proposta por Jacobs *et al.* (1991) e utilizada por Melo *et al.* (2003) e Pinto (2005) as etapas de partição do espaço de entrada e de estimação dos parâmetros dos especialistas locais são isoladas. Neste trabalho, a integração das etapas de partição do espaço de entrada e de estimação dos parâmetros dos especialistas locais apresenta-se como uma possibilidade de melhoria tanto na qualidade do ajuste, na fase de estimação dos parâmetros, como na utilização dos modelos para fazer previsões.

Como foi apresentado anteriormente, tanto o algoritmo k-médias como o método dos mínimos quadrados podem ser vistos como modelos de otimização, nos quais pretende-se otimizar o valor de uma função objetivo respeitando um conjunto de restrições. Da mesma forma, para fazer a integração das etapas de partição do espaço de entrada e de estimação dos parâmetros dos especialistas locais, em problemas de regressão, propõe-se uma formulação que incorpore a partição do espaço de entrada na estimação dos parâmetros dos modelos de regressão locais. Assim, se for tomada como função de perda, que deve ser minimizada, a de mínimos quadrados, chega-se no seguinte modelo de otimização

$$Min \ \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \sum_{c=1}^{k} P_{ic} \cdot f(x_i)_c \right]^2 \tag{9}$$

em que

$$P_{ic} = \frac{e^{-\left(\sum_{j=1}^{p}(x_{ij}-m_{cj})^2\right)^{1/2}}}{\sum_{c=1}^{k} e^{-\left(\sum_{j=1}^{p}(x_{ij}-m_{cj})^2\right)^{1/2}}} \qquad (10)$$

no caso em que o grau de pertinência das observações aos agrupamentos (P_{ic}) é dada pela função *softmax*. No caso linear, $f(X_i)_c = \alpha_c + \beta_c.x_i$ e, no caso quadrático, $f(X_i)_c = \beta_{0c} + \beta_{1c}.x_i + \beta_{2c}.x_i^2$, em que α_c e β_c (c=1,...,k) são os parâmetros a serem estimados no caso linear e β_{0c}, β_{1c} e β_{2c} (c=1,...,k) são os parâmetros a serem estimados no caso quadrático. Este modelo de otimização é não-linear irrestrito, a função de perda é a de mínimos quadrados e as variáveis de decisão são os centróides dos agrupamentos (m_{cj}) e os parâmetros dos especialistas.

Variações desta formulação podem ser obtidas modificando-se a função de perda ou a função que determina o grau de pertinência das observações aos agrupamentos.

3.1 Algoritmo para resolução do problema de otimização proposto

Para a resolução deste modelo de otimização não-linear irrestrito uma alternativa é utilizar um time assíncrono (Saito Jr., 1999) que é uma estrutura que utiliza algoritmos com características diferentes cooperando entre si, o que aumenta a chance do treinamento obtido ser mais eficiente e que gera um resultado melhor do que quando cada um dos algoritmos trabalha de forma independente.

Assim, a opção é utilizar um time assíncrono composto por uma meta-heurística evolutiva baseada em algoritmos genéticos e pelo método Quasi-Newtoniano. O papel da meta-heurística evolutiva é gerar uma solução inicial que é utilizada como ponto de partida para o método Quase-Newtoniano.

Segundo Saito Jr. (1999) a justificativa de utilizar uma estrutura de algoritmos trabalhando de modo paralelo e assíncrono é que alguns dos algoritmos de otimização existentes podem apresentar dificuldades para conseguir obter o ponto mínimo global de uma função objetivo, em um problema de minimização. Um algoritmo, para resolver problemas de otimização, pode apresentar pontos fracos como a convergência para mínimos locais, o que ocorre pois muitos desses algoritmos são do tipo *hill-climbing*. Essa característica não é desejada já que não é interessante terminar o processo de treinamento em um ponto de mínimo local, em que o valor da função objetivo ainda é considerado alto (Saito Jr., 1999).

Segundo Goldberg (1989) uma meta-heurística evolutiva lida com uma população de soluções, que evolui, principalmente, através da iteração entre seus elementos. Segundo Konzen *et al.* (2003) o algoritmo genético é uma meta-heurística evolutiva, sendo uma técnica de busca aleatória direcionada, desenvolvida por Holland (1975), capaz de obter a solução ótima global em um espaço de busca complexo multi-dimensional. Seu princípio de funcionamento é baseado no processo evolutivo dos seres vivos, seguindo o princípio básico de que as gerações derivadas serão mais evoluídas do que seus precursores (Cooper, 2000).

De acordo com Konzen *et al.* (2003) a evolução da população é realizada através de operadores genéticos denominados *crossover* e mutação. No *crossover* os cromossomos são

combinados formando novos indivíduos e na mutação, os componentes de uma solução podem sofrer perturbações em seus genes.

Ainda, segundo Konzen *et al.* (2003), pode-se descrever o funcionamento dos algoritmos genéticos nos passos:

Passo 1 – Fazer a codificação das variáveis. Tradicionalmente, a codificação é feita criando-se códigos binários (Goldberg, 1989). Embora essa forma de representação tenha se mostrado eficiente em vários problemas, observou-se, a medida que foram crescendo as aplicações de algoritmos genéticos, que esta representação pode não ser a mais adequada, surgindo daí alternativas como a representação por números inteiros ou reais, em que o cromossomo é descrito por um vetor desses números (Claumann, 1999). Ainda, segundo Claumann (1999) as vantagens da codificação real em relação à codificação binária são:

a) Na codificação real não há necessidade de conversões para avaliar a função objetivo, pois cada gene corresponde a uma variável. Em codificação binária, vários genes são utilizados para representar uma única variável;
b) O tempo de processamento tende a ser menor na codificação real, em problemas multivariáveis, pois a codificação binária pode gerar cromossomos muito grandes, sendo que, a aplicação dos operadores genéticos demandará um elevado tempo computacional;
c) O limite de precisão da solução obtida em codificação real é o da precisão da máquina. Em codificação binária este limite é baseado no número de genes (*bits*) utilizados na representação das variáveis;
d) A utilização da codificação real permite um maior controle em relação à ação dos operadores genéticos, pois cada gene representa uma variável. No caso da codificação binária, a aplicação dos operadores genéticos produz modificações nos fenótipos que são difíceis de serem previstas.

Recentemente, tem-se utilizado a codificação em números reais, ao invés de codificações binárias, principalmente, para resolver problemas de otimização na formação de agrupamentos (Niesse & Mayne, 1996 e Iwamatsu, 2000), o que possibilitou uma melhora significativa nos resultados obtidos, quando comparado à codificação binária.

Passo 2 – Gerar, aleatoriamente, uma população inicial de soluções. Cada solução (cromossomos ou indivíduos) é formada por uma cadeia de variáveis de decisão;

Passo 3 – Avaliar o *fitness* dos indivíduos da população. Uma forma de avaliar a solução é calculando o valor da função objetivo da mesma;

Passo 4 – Repetir até que o critério de parada seja atendido

a) Selecionar um conjunto de pais na população e realizar a mutação;
b) Cruzar os pais de modo que se reproduzam (*crossover*);
c) Avaliar a *fitness* dos filhos gerados;
d) Substituir os filhos julgados inadequados.

Segundo Iwamatsu (2000) quando a codificação empregada é em números reais, diferentes operadores genéticos podem ser utilizados para que o Passo 4 seja mais eficiente como: inversão, média aritmética, média geométrica e *crossover* m-pontos. Os detalhes desses operadores genéticos podem ser encontrados em Niesse & Mayne (1996). Além do *crossover*

m-pontos, fêz-se uso do *crossover* uniforme, que combina, aleatoriamente, as varáveis de decisão dos cromossomos pais, para aumentar a eficiência do algoritmo.

Passo 5 – Fim

Em relação à seleção dos pais da população, normalmente, estes são ordenados de acordo com o valor da *fitness* e divididos em grupos. O grupo com as piores soluções é descartado e o grupo com as melhores soluções é mantido e sofre mutação e *crossover* até que todos os indivíduos descartados sejam repostos.

Além da meta-heurística evolutiva baseada em algoritmos genéticos, o time assíncrono é composto pelo método Quase-Newtoniano.

Segundo Bazaraa *et al.* (1993), o método Quasi-Newtoniano foi originalmente proposto por Davidon (1959) e, posteriormente, desenvolvido por Fletcher & Powell (1963). Uma característica desse método é que utiliza uma aproximação da matriz Hessiana da função objetivo, para guiar a busca pela solução ótima, ao invés da própria matriz Hessiana, uma vez que a sua estimação, principalmente das derivadas segundas, pode ser difícil de se obter.

Segundo Fletcher & Powell (1963), pelo método Quasi-Newtoniano, a cada passo, avalia-se o valor da função objetivo em diferentes pontos, para estimar a derivada primeira e a derivada segunda da mesma. Esses valores permitem decidir a direção a ser seguida para atingir o ótimo da função.

3.2 Ilustração do modelo de otimização proposto e do algoritmo de resolução

Para ilustrar o funcionamento do time assíncrono na abordagem de composição de especialistas com otimização na formação dos agrupamentos, em problemas de regressão, gerou-se um conjunto de treinamento com 30 observações, no espaço 1-dimensional, em que a variável dependente foi obtida pela composição de dois especialistas lineares. Em relação à função que gerou o grau de pertinência das observações aos agrupamentos foi utilizada a função *softmax*. Um ruído aleatório foi incorporado à série gerada. A Tabela 1 mostra os parâmetros dos especialistas e os centróides dos agrupamentos utilizados na geração das observações, a Tabela 2 mostra os dados gerados e a Figura 1 o conjunto de treinamento gerado.

Tabela 1 – Parâmetros dos especialistas e centróides dos agrupamentos.

Parâmetro	Valor
Centróide do agrupamento 1	$m_1 = 11{,}200$
Centróide do agrupamento 2	$m_2 = 19{,}700$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 3{,}600$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -0{,}730$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -3{,}300$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}450$
Média do ruído	0,000
Desvio padrão do ruído	0,500

Tabela 2 – Observações geradas pela composição de dois especialistas lineares.

X	**Y**	X	**Y**	X	**Y**
8,0	**-2,875**	13,0	**-5,681**	18,0	**5,233**
8,5	**-2,066**	13,5	**-6,468**	18,5	**5,132**
9,0	**-2,519**	14,0	**-6,784**	19,0	**5,931**
9,5	**-3,326**	14,5	**-5,635**	19,5	**5,074**
10,0	**-3,688**	15,0	**-4,452**	20,0	**5,091**
10,5	**-4,048**	15,5	**-1,833**	20,5	**5,568**
11,0	**-4,713**	16,0	**1,164**	21,0	**5,912**
11,5	**-4,574**	16,5	**2,463**	21,5	**6,241**
12,0	**-4,807**	17,0	**3,877**	22,0	**6,739**
12,5	**-5,847**	17,5	**4,031**		

O funcionamento do time assíncrono, composto por uma meta-heurística evolutiva, baseada em algoritmos genéticos e no método Quasi-Newtoniano, para a estimação dos centróides e dos parâmetros dos modelos, obedece aos seguintes passos:

Passo 1: Gerar uma população inicial de soluções

Seguindo a sugestão de Liu & Xie (1995) em relação ao tamanho da população de soluções, gerou-se um conjunto de 100 soluções iniciais, aleatoriamente. Empregou-se a codificação em números reais e a representação usada foi uma cadeia de variáveis de decisão com 6 elementos, em que o primeiro dos valores é o centróide do agrupamento 1, o segundo é o centróide do agrupamento 2, o terceiro é o intercepto do modelo linear para o agrupamento 1, o quarto é o coeficiente angular do modelo linear para o agrupamento 1, o quinto é o intercepto do modelo linear para o agrupamento 2 e o sexto valor é o coeficiente angular do modelo linear para o agrupamento 2. Em relação aos parâmetros utilizados na geração aleatória das soluções, para os centróides, os valores são uniformemente distribuídos entre 10 e 20, para os interceptos e para os coeficientes angulares, os valores são uniformemente distribuídos entre -5 e 5.

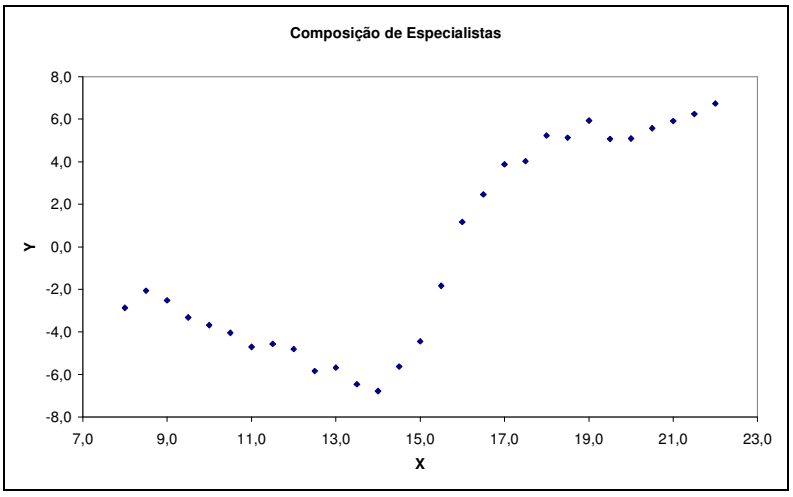

Figura 1 – Observações geradas pela composição dos especialistas.

Passo 2: Avaliar as soluções geradas

Nessa etapa calcula-se o valor da função objetivo, para as 100 soluções iniciais geradas. Essas soluções foram, então, classificadas em ordem crescente.

Passo 3: Seleção dos pais e cruzamentos

Nesta etapa as 10 soluções com menor valor da função objetivo são mantidas e as 90 restantes são excluídas. A partir das 10 soluções selecionadas são feitos cruzamentos para se obter 90 novas soluções. Neste trabalho, os cruzamentos utilizados são apenas do tipo *crossover* uniforme, ou seja, optou-se por não utilizar outro operador genético. Optou-se, também, por não realizar mutação.

Passo 4: Convergência da meta-heurística evolutiva baseada em algoritmos genéticos

Se o algoritmo não convergiu volta-se para o passo 2, caso contrário segue-se para o passo 5. O critério de convergência adotado foi de 0,01.

Passo 5: Refinamento da solução aplicando-se o método Quasi-Newtoniano, tomando como valor inicial a solução obtida pela meta-heurística evolutiva, baseada em algoritmos genéticos.

Os parâmetros foram estimados em um microcomputador com processador Celeron 600 MHz e 192 MB de memória RAM, utilizando o programa computacional SAS versão 8.2. O tempo demandado na estimação dos parâmetros pelo time assíncrono composto pela meta-heurística evolutiva, baseada em algoritmos genéticos, e pelo método Quasi-Newtoniano foi de 4 segundos e 92 centésimos.

A Tabela 3 mostra os parâmetros dos especialistas e os centróides dos agrupamentos estimados pela meta-heurística evolutiva baseada em algoritmos genéticos, a Figura 2 mostra a aderência obtida pelos parâmetros e centróides estimados pela meta-heurística evolutiva baseada em algoritmos genéticos, a Tabela 4 mostra os parâmetros dos especialistas e os centróides dos agrupamentos finais obtidos e a Figura 3 mostra a aderência final obtida.

Verifica-se, a partir das Tabelas 3 e 4 e Figuras 2 e 3, que a meta-heurística evolutiva baseada em algoritmos genéticos gerou uma boa aproximação da solução inicial e que o método Quasi-Newtoniano, a partir dessa solução, chegou em uma solução final muito próxima dos parâmetros reais. Em relação ao valor da função objetivo, no passo 1, a solução com menor valor de função objetiva, obtida de forma aleatória, foi de 16,415. Após a convergência da meta-heurística evolutiva baseada em algoritmos genéticos esse valor foi para 0,971 e a solução final obtida teve o valor da função objetivo igual a 0,115.

A Tabela 5 compara os valores reais dos parâmetros utilizados na geração da composição de especialistas e estimados pelo time assíncrono, assim como o valor de suas funções objetivo.

Tabela 3 – Parâmetros estimados pela meta-heurística evolutiva.

Parâmetro	Valor
Centróide do agrupamento 1	$m_1 = 11{,}466$
Centróide do agrupamento 2	$m_2 = 19{,}335$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 0{,}283$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -0{,}428$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -1{,}648$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}300$
Valor da função objetivo	FO = 0,971

Tabela 4 – Estimativas finais obtidas.

Parâmetro	Valor
Centróide do agrupamento 1	$m_1 = 11{,}210$
Centróide do agrupamento 2	$m_2 = 19{,}710$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 3{,}616$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -0{,}736$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -3{,}295$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}447$
Valor da função objetivo	FO = 0,124

Tabela 5 – Comparação dos parâmetros utilizados e estimados pelo algoritmo.

Parâmetro	Valor utilizado	Valor estimado
Centróide do agrupamento 1	$m_1 = 11{,}200$	$m_1 = 11{,}210$
Centróide do agrupamento 2	$m_2 = 19{,}700$	$m_2 = 19{,}710$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 3{,}600$	$\alpha_1 = 3{,}616$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -0{,}730$	$\beta_1 = -0{,}736$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -3{,}300$	$\alpha_2 = -3{,}295$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}450$	$\beta_2 = 0{,}447$
Valor da função objetivo	FO = 0,128	FO = 0,124

Verifica-se na Tabela 5 que o valor dos parâmetros obtidos utilizando o time assíncrono são próximos dos valores utilizados na geração da composição de especialistas e que o valor da função objetivo, calculada a partir dos parâmetros estimados, é menor que o valor da função objetivo, calculada a partir dos parâmetros utilizados na geração da composição. Desta forma, é possível afirmar que o time assíncrono foi capaz de obter uma boa solução para problema de otimização não-linear irrestrito proposto.

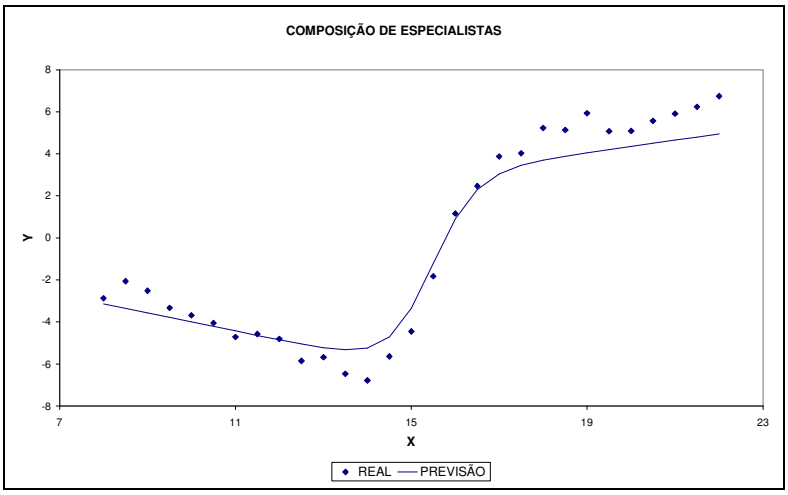

Figura 2 – Aderência obtida a partir dos parâmetros estimados pela meta-heurística evolutiva.

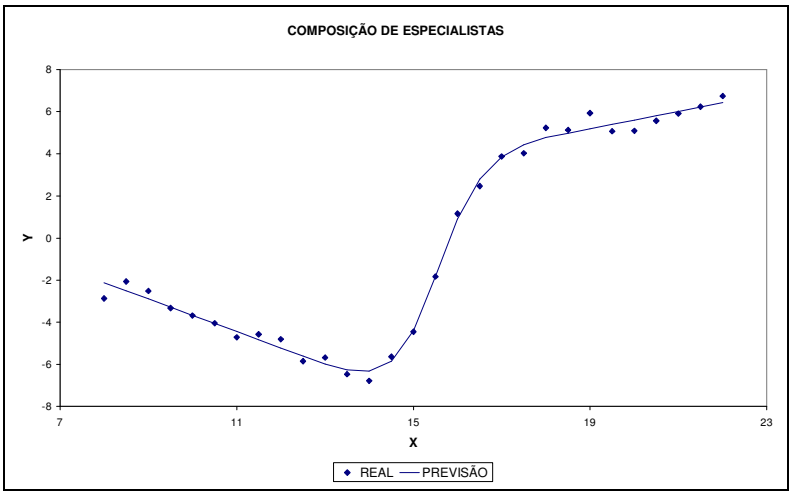

Figura 3 – Aderência final obtida.

3.3 Estimação dos centróides e dos parâmetros do modelo pelo método seqüencial em grade regular

Empregou-se uma estratégia de busca sistemática denominada método seqüencial em grade regular para que se tivesse um parâmetro de comparação do desempenho do time assíncrono na estimação dos centróides e dos parâmetros dos modelos. Os valores utilizados na geração das soluções, para os centróides dos agrupamentos, variaram de 10 a 20 com espaçamento 1, ou seja, para o centróide de cada um dos agrupamentos os valores foram 10, 11, 12, ..., 18, 19 e 20. Para os interceptos e para os coeficientes angulares, os valores variaram de -5 a 5 com espaçamento 1, ou seja, foram -5, -4, -3, ..., 3, 4 e 5. A combinação desses valores totalizou 1.771.561 ($=11^6$) soluções e para cada uma dessas soluções calculou-se o valor da função objetivo dada por

$$\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[y_i - \sum_{c=1}^{k} \frac{e^{-\left(\sum_{\mathrm{j}=1}^{\mathrm{p}} (x_{ij}-m_{cj})^2\right)^{1/2}}}{\sum_{c=1}^{k} e^{-\left(\sum_{\mathrm{j}=1}^{\mathrm{p}} (x_{ij}-m_{cj})^2\right)^{1/2}}} (\alpha_c + \beta_c . x_i) \right]^2 \qquad (11)$$

em que α_c e β_c são os parâmetros e m_{cj} são os centróides dos agrupamentos a serem estimados.

Posteriormente, as soluções foram ordenadas de forma crescente, em relação ao valor da função objetivo, e a solução com o menor erro foi selecionada.

Os parâmetros foram estimados em um microcomputador com processador Celeron 600 MHz e 192 MB de memória RAM, utilizando o programa computacional SAS versão 8.2. O tempo demandado na estimação dos parâmetros pelo método exaustivo foi de 23 minutos 9 segundos e 88 centésimos.

A Tabela 6 mostra os parâmetros dos especialistas, os centróides dos agrupamentos e o valor da função objetivo reais e os estimados pelo método seqüencial em grade regular e a Figura 4 mostra a aderência obtida a partir deste método de estimação.

Verifica-se pela Tabela 6 que há uma grande diferença entre os valores reais e os obtidos pelo método seqüencial em grade regular. Para explicar porque a solução obtida pelo método seqüencial em grade regular é muito diferente dos valores reais, calculou-se o valor da função objetivo obtida arredondando-se os valores reais. A Tabela 7 mostra os parâmetros dos especialistas, os centróides dos agrupamentos e o valor da função objetivo estimados pelo método seqüencial em grade regular e pelos valores reais arredondados.

Tabela 6 – Parâmetros estimados pelo método seqüencial em grade regular e valores reais.

Parâmetro	Valor real	Valor estimado
Centróide do agrupamento 1	$m_1 = 11{,}200$	$m_1 = 15{,}000$
Centróide do agrupamento 2	$m_2 = 19{,}700$	$m_2 = 16{,}000$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 3{,}600$	$\alpha_1 = 2{,}000$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -0{,}730$	$\beta_1 = -1{,}000$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -3{,}300$	$\alpha_2 = -5{,}000$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}450$	$\beta_2 = 1{,}000$
Valor da função objetivo	FO = 0,128	FO = 1,193

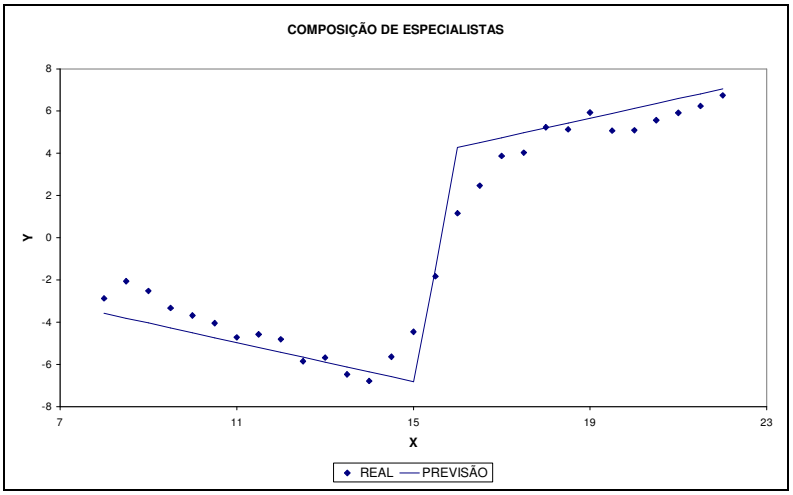

Figura 4 – Aderência obtida pelos parâmetros estimados pelo método seqüencial em grade regular.

Verifica-se pela Tabela 7 que o valor da função objetivo obtido arredondando-se os valores reais é muito mais alto que o obtido pelo método seqüencial em grade regular, o que torna a solução obtida aplicando-se o método seqüencial em grade regular mais adequada. Essa grande diferença pode ser explicada pelo espaçamento utilizado na geração das soluções do método seqüencial em grade regular. Assim, resultados mais próximos podem ser obtidos reduzindo-se o espaçamento, porém haveria um número muito maior de soluções, o que demandaria mais tempo de processamento.

Tabela 7 – Parâmetros estimados pelo método seqüencial em grade regular e valores reais arredondados.

Parâmetro	Valor real arredondado	Valor estimado
Centróide do agrupamento 1	$m_1 = 11{,}000$	$m_1 = 15{,}000$
Centróide do agrupamento 2	$m_2 = 20{,}000$	$m_2 = 16{,}000$
Intercepto do especialista do agrupamento 1	$\alpha_1 = 4{,}000$	$\alpha_1 = 2{,}000$
Coeficiente angular do especialista do agrupamento 1	$\beta_1 = -1{,}000$	$\beta_1 = -1{,}000$
Intercepto do especialista do agrupamento 2	$\alpha_2 = -3{,}000$	$\alpha_2 = -5{,}000$
Coeficiente angular do especialista do agrupamento 2	$\beta_2 = 0{,}000$	$\beta_2 = 1{,}000$
Valor da função objetivo	FO = 35,340	FO = 1,193

4. Utilização da abordagem proposta na previsão da receita líquida das empresas de transporte aéreo

As abordagens especialista global e composição de especialistas locais com otimização na formação dos agrupamentos foram aplicadas na previsão da receita líquida (R$ mil de 2002) das empresas de transporte aéreo.

Utilizou-se como variável explicativa o ativo total das empresas de transporte aéreo (R$ mil de 2002). Os dados utilizados são provenientes das revistas Balanço Anual dos anos de 2002 e de 2003 e do anuário do DAC (Departamento de aviação civil) do ano de 2004 totalizando 39 observações. Essas observações foram divididas, de forma aleatória, em 2 grupos. O primeiro grupo ficou com 26 observações e foi utilizado no treinamento dos modelos (conjunto de treinamento) e o segundo grupo ficou com 13 observações e foi utilizado na validação dos modelos (conjunto de validação). Tirou-se o logaritmo da receita líquida e do ativo total para que o estudo contemplasse empresas de diferentes portes.

Para a escolha dos especialistas utilizou-se como indicadores, o erro absoluto percentual médio (EAPM) e a média da soma dos erros quadráticos (MEQ) calculados por:

$$EAPM = \frac{1}{N}\sum_{t=1}^{N}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t} \qquad MEQ = \frac{1}{N}\sum_{t=1}^{N}\left(Y_t - \hat{Y}_t\right)^2$$

em que N é número de observações, Y_t é o logaritmo da receita líquida das empresas de transporte aéreo no período t e $\hat{Y}_t$ é o logaritmo da receita líquida prevista das empresas de transporte aéreo, calculados a partir do conjunto de validação. Foram eleitos os especialistas com o menor valor desses indicadores.

4.1 Especialista global

Um especialista é chamado de global quando é obtido a partir de todos os dados do conjunto de treinamento. Neste estudo de caso, os modelos linear e quadrático foram treinados e testados para selecionar o especialista global. As estatísticas de desempenho geradas pela aplicação dos especialistas no conjunto de validação mostram que o modelo linear foi o de melhor desempenho na fase de validação, uma vez que apresenta o menor valor de EAPM

(5,05% contra 6,02% para o modelo quadrático) e a menor MEQ (0,394 contra 0,533 para o modelo quadrático). Assim, como especialista global o modelo selecionado é o linear dado por

$$RL = + 1{,}0167.AT - 0{,}1740 \qquad (R^2 = 0{,}858)$$

em que RL é o logarítmo da receita líquida prevista e AT é o logaritmo do ativo total. As Figuras 5 e 6 mostram a aderência do especialista global (modelo linear) nos conjuntos de treinamento e de validação, respectivamente.

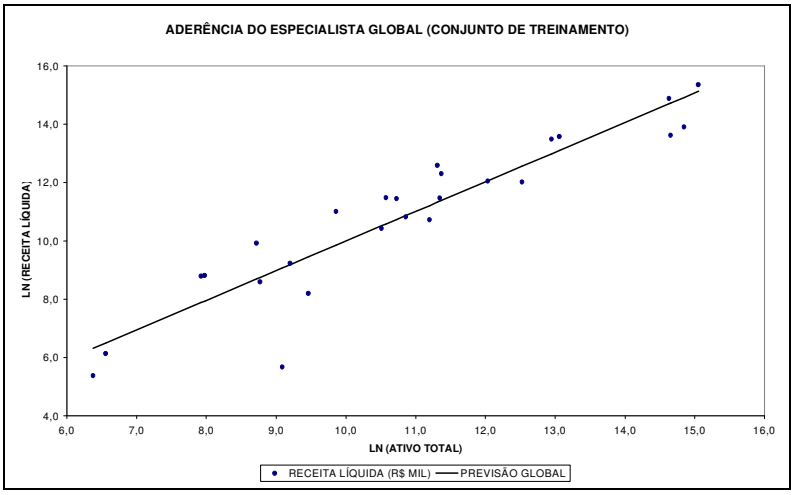

Figura 5 – Aderência do especialista global no conjunto de treinamento.

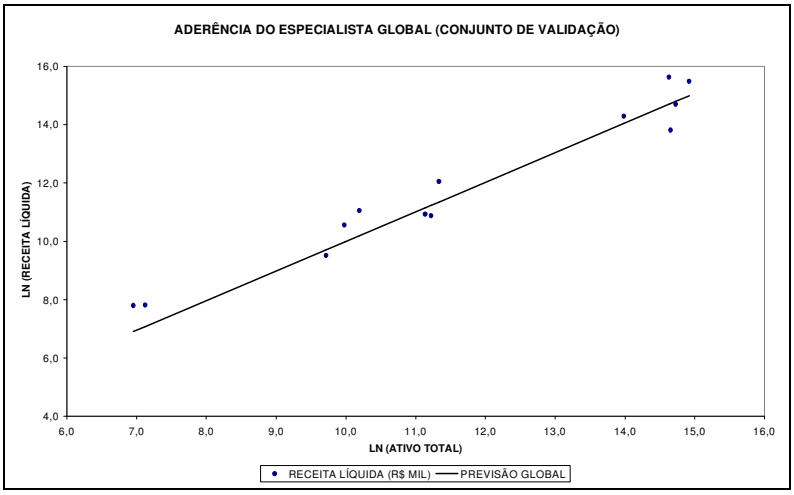

Figura 6 – Aderência do especialista global no conjunto de validação.

4.2 Composição de especialistas com otimização na formação dos agrupamentos

Como a otimização na formação dos agrupamentos, pela abordagem proposta, é feita de forma integrada, há a necessidade de se aplicar a metodologia em todas as possíveis combinações de composição de especialistas. Como neste estudo de caso considerou-se apenas 2 agrupamentos e 2 tipos de especialista (modelo linear e modelo quadrático), existem apenas 3 possibilidades de composição de especialistas: 1) nos dois agrupamentos o modelo vencedor ser o linear, que foi denominada composição linear-linear; 2) em um dos agrupamentos o modelo vencedor ser o linear e no outro o modelo vencedor ser o quadrático, que foi denominada composição linear-quadrático e 3) nos dois agrupamentos o modelo vencedor ser o quadrático, que foi denominada composição quadrático-quadrático. Como função que determina a pertinência das observações aos agrupamentos utilizou-se a *softmax*.

Desta forma, na composição linear-linear (LL) tem-se o modelo de otimização:

$$Min\ \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[y_i-\left(P_{i1}\left(\alpha_1+\beta_1.AT_i\right)+P_{i2}\left(\alpha_2+\beta_2.AT_i\right)\right)\right]^2 \tag{12}$$

Na composição linear-quadrático (LQ) tem-se o modelo de otimização:

$$Min\ \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[y_i-\left(P_{i1}\left(\alpha_1+\beta_1.AT_i\right)+P_{i2}\left(\alpha_2+\beta_2.AT_i+\delta_2.AT_i^2\right)\right)\right]^2 \tag{13}$$

E na composição quadrático-quadrático (QQ) tem-se o modelo de otimização:

$$Min\ \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[y_i-\left(P_{i1}\left(\alpha_1+\beta_1.AT_i+\delta_1.AT_i^2\right)+P_{i2}\left(\alpha_2+\beta_2.AT_i+\delta_2.AT_i^2\right)\right)\right]^2 \tag{14}$$

em que

$$P_{i1}=\frac{e^{-\left((AT_i-m_1)^2\right)^{1/2}}}{e^{-\left((AT_i-m_1)^2\right)^{1/2}}+e^{-\left((AT_i-m_2)^2\right)^{1/2}}} \tag{15}$$

e

$$P_{i2}=\frac{e^{-\left((AT_i-m_2)^2\right)^{1/2}}}{e^{-\left((AT_i-m_1)^2\right)^{1/2}}+e^{-\left((AT_i-m_2)^2\right)^{1/2}}} \tag{16}$$

sendo α_1, β_1, δ_1, α_2, β_2, δ_2, m_1 e m_2 os parâmetros a serem estimados.

As estatísticas de desempenho geradas pelas diferentes composições, no conjunto de validação, mostram que a composição linear-quadrático foi a com melhor desempenho, uma vez que apresentou o menor valor de EAPM (4,68% contra 4,73% para a composição quadrático-quadrático e 5,36% para a composição linear-linear) e a menor SEQ (0,346 contra 0,349 para a composição quadrático-quadrático e 0,466 para a composição linear-linear). Desta forma, para se fazer previsão utilizando a composição de especialistas vencedora utiliza-se:

$$RL_i = P_1.(4,725 + 0,656.AT_i) + P_2.(-75,907 + 21,327.AT_i - 1,345.AT_i^2)$$

em que

$$P_1 = \frac{e^{-\left((AT_i - 15,759)^2\right)^{1/2}}}{e^{-\left((AT_i - 15,759)^2\right)^{1/2}} + e^{-\left((AT_i - 2,639)^2\right)^{1/2}}}$$

e

$$P_2 = \frac{e^{-\left((AT_i - 2,639)^2\right)^{1/2}}}{e^{-\left((AT_i - 15,759)^2\right)^{1/2}} + e^{-\left((AT_i - 2,639)^2\right)^{1/2}}}$$

sendo RL_i a previsão da receita líquida logaritmizada da empresa i e AT_i o ativo total logaritmizado da empresa i. As Figuras 7 e 8 mostram a aderência da composição de especialistas com otimização na formação dos agrupamentos vencedora nos conjuntos de treinamento e de validação, respectivamente.

O desempenho das abordagens especialista global e composição de especialistas locais com otimização na formação de agrupamentos foi comparado em relação às previsões geradas (conjunto de validação). As estatísticas de desempenho mostram que a abordagem com o melhor desempenho foi a de composição de especialistas com a otimização da formação de agrupamentos. Adotando-se a estatística de desempenho EAPM, o uso desta abordagem reduziu em 7,3% o erro de previsão, quando comparado ao especialista global. Já adotando-se a estatística de desempenho MEQ, o uso desta abordagem reduziu em 12,2% o erro de previsão, quando comparado ao especialista global.

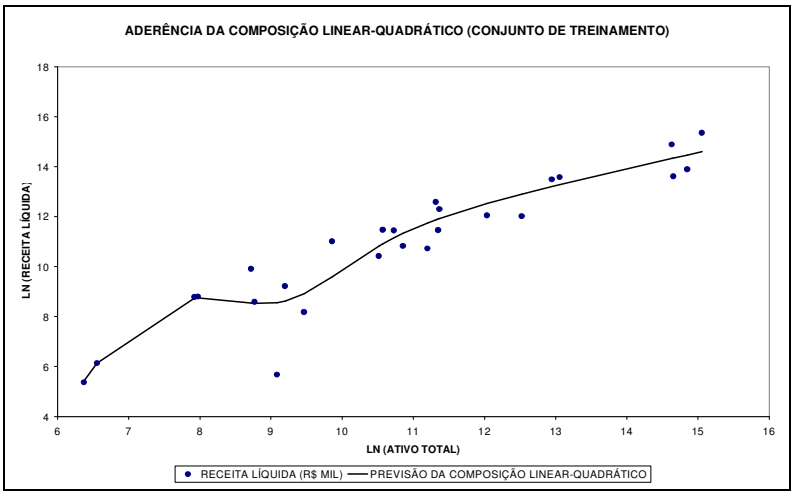

Figura 7 – Aderência da composição de especialistas locais com otimização na formação dos agrupamentos no conjunto de treinamento.

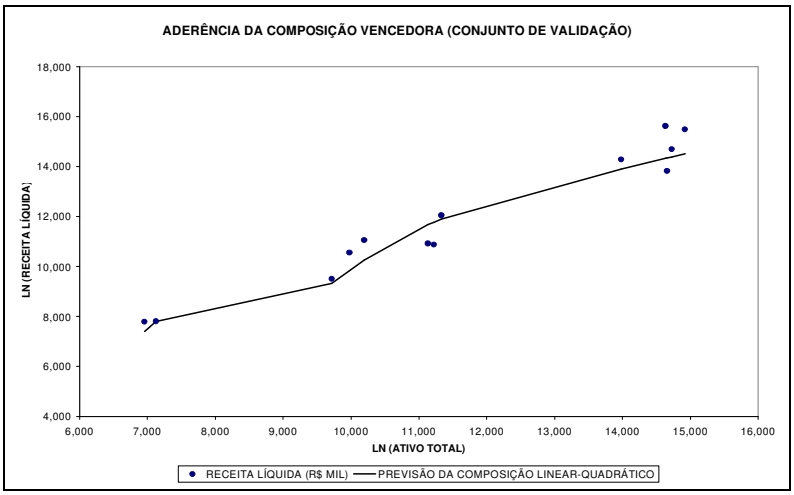

Figura 8 – Aderência da composição de especialistas locais com otimização na formação dos agrupamentos no conjunto de validação.

5. Conclusão

A composição de especialistas locais tem uso potencial em uma grande variedade de problemas referentes a modelagem, principalmente, quando o objetivo é fazer previsões, que é uma área fundamental das análises e tomadas de decisão e que, até hoje, continua sendo um desafio para os pesquisadores.

Neste trabalho foi proposta a integração das etapas de formação de agrupamentos e designação dos especialistas, em composição de especialistas locais, aplicados a problemas de regressão visando melhorar a qualidade dos ajustes dos modelos, melhorar a qualidade das previsões realizadas e otimizar a formação dos agrupamentos. Para cumprir com esses objetivos foi proposta uma formulação que integrasse a estimação dos parâmetros que definem os especialistas e os agrupamentos.

Como proposta para futuras pesquisas pretende-se estender a abordagem de composição de especialistas com otimização na formação dos agrupamentos para problemas de classificação.

Agradecimentos

Este trabalho foi financiado pela FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo.

Referências Bibliográficas

(1) Bazaraa, M.S.; Sherali, H.D. & Shetty, C.M. (1993). *Nonlinear programming: theory and applications*. 2ª edition, John Wiley & Sons, Inc.

(2) Claumann, C.A. (1999). Modelagem dinâmica e controle de processos não lineares: uma aplicação de algoritmos genéticos para treinamento de redes neurais recorrentes. Dissertação de mestrado em Engenharia Química, Programa de Pós-graduação em Engenharia Química, UFSC, Florianópolis, SC, Brasil.

(3) Cooper, B. (2000). Modelling research and development: How do firms solve design problems? *Journal of Evolutionary Economics*, **10**, 395-413.

(4) Davidon, W.C. (1959). Variable metric method for minimization. AEC Research Development Report, ANL-5990.

(5) Duda R.O.; Hart, P.E. & Stork, D.G. (2001). *Pattern Classification*. 2ª edition, John Wiley & Sons, Inc. New York.

(6) Fletcher, R. & Powell, M. (1963). A rapidly convergent descent method for minimization. *Computer Journal*, **6**, 163-168.

(7) Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

(8) Holland, J.H. (1975). *Adaptation in natural and artificial systems*. MIT Press, Ann Arbor, Michigan.

(9) Huber, P. (1964). Robust estimation of a location parameter. *Annals on Mathematical Statistics*, **35**, 73-101.

(10) Iwamatsu, M. (2000). Global geometry optimization of silicon clusters using the space-fixed genetic algorithm. *Journal of Chemical Physics*, **112**(24), 10976-10983.

(11) Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J. & Hinton, G.E. (1991). Adaptive mixture of local experts. *Neural Computation*, **3**(1), 79-87, MIT Press.

(12) Konzen, P.H.A.; Furtado, J.C.; Carvalho, C.W.; Ferrão, M.F.; Molz, R.F.; Bassani, I.A. & Hüning, S.L. (2003). Otimização de métodos de controle de qualidade de fármacos usando algoritmo genético e busca tabu. *Pesquisa Operacional*, **23**(1), 189-207.

(13) Liu, J. & Xie, W. (1995). A Genetics-Based Approach to Fuzzy Clustering. *Proceedings of the IEEE International Conference on Fuzzy Systems*.

(14) Mangasarian, O.L. (1997). Mathematical Programming in Data Mining. *Data Mining and Knowledge Discovery*, **1**(2), 183-201.

(15) Melo, B.; Milioni, A.Z. & Nascimento Jr., C.L. (2003). Modelos de especialistas globais e de composição de especialistas locais para previsão de séries temporais. **In**: *XXXV SBPO – Simpósio Brasileiro de Pesquisa Operacional*.

(16) Niesse, J. & Mayne, H. (1996). Global geometry optimization of atomic clusters using a modified genetic algorithm in space-fixed coordinates. *Journal of Chemical Physics*, **105**(11), 4700-4706.

(17) Pinto, D.B.T. (2003). MLEM on cross section data. Trabalho de graduação, Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP.

(18) Pinto, D.B.T. (2005). Choices and pitfalls concerning mixture-of-experts modeling. Tese (Mestrado em Engenharia Aeronáutica e Mecânica, área de Produção), Instituto Tecnológico de Aeronáutica, São José dos Campos.

(19) Saito Jr., P.A. (1999). Treinamento de redes neurais utilizando time assíncrono. Tese (Mestrado em Engenharia Eletrônica e Computação, área de Sistemas e Controle), Instituto Tecnológico de Aeronáutica, São José dos Campos.

(20) Vapnik, V.N. (1995). *The nature of statistical learning theory*. Springer, New York.

(21) Vapnik, V.N. (1999). An overview of Statistical learning theory. *IEEE transactions on neural networks*, **10**(5), 988-999.

(22) Webb, A. (2002). *Statistical Pattern Recognition*. 2ª edition, John Wiley & Sons, Inc.