COMPACTAÇÃO DE DADOS DE CROMATOGRAFIA GASOSA PARA CLASSIFICAÇÃO NEURAL DE ÓLEOS

Rosa Cristhyna de Oliveira Vieira Paes[1]
rosa.paes@lab2m.coppe.ufrj.br

Ana Cristina da Silva Serra[2]
ana.serra@lab2m.coppe.ufrj.br

[1, 2] Universidade Federal do Rio de Janeiro, LAB2M/COPPE/UFRJ - Bloco I-2000/Anexo, Ilha do Fundão, 21949-900, Rio de Janeiro, RJ, Brasil

ABSTRACT

Nowadays, since several sedimentary basins have been extensively studied, using geochemical analysis, to find a computational tool that make possible to classify different types of oil fast and accurately would become more agile the characterization of petroleum systems. Since there are a plenty of data to be analyzed, the aim is to become the process automated, to generate accurate results as fast as possible, that would reduce the duration of oil analyses. Firstly the neural network technique was applied and its performance was evaluated to classify different types of oil samples, according to the *n*-alkanes distribution profiles in gas chromatograms. The algorithm obtained provided a classification of oils with a high level of accuracy. Once a network has been trained, the objective of this study was to evaluate its performance when the data are compacted through Principal Components Analysis (PCA) and variables Relevance Analysis.

KEYWORDS: gas chromatograms; classification; neural network; principal components analysis; relevance analysis

RESUMO

Atualmente, devido ao fato de diversas bacias sedimentares serem alvo de estudos geoquímicos detalhados, a obtenção de uma ferramenta computacional que permita classificar diferentes tipos de óleos de maneira rápida e precisa possibilitaria uma agilização para caracterização de sistemas petrolíferos. Em virtude da grande quantidade de dados a ser analisada, procura-se tornar o processo o mais automatizado possível, com geração de resultados rápidos e precisos, cujos benefícios concentram-se na redução do tempo de análise do tipo de óleo. Inicialmente aplicou-se a técnica de redes neurais e seu desempenho foi avaliado para classificação de diferentes tipos de óleo, de acordo com os perfis de distribuição dos *n*-alcanos em cromatogramas gasosos. O algoritmo obtido permitiu uma classificação de diferentes tipos de óleos com um alto nível de confiabilidade. A partir da rede treinada, o objetivo desse estudo consistiu em avaliar seu desempenho quando os dados são compactados através das técnicas de Análise de Componentes Principais (PCA) e Análise de Relevância das variáveis.

PALAVRAS-CHAVE: cromatogramas gasosos; classificação; redes neurais; análise de componentes principais; relevância

1 INTRODUÇÃO

Nos últimos anos diversas bacias sedimentares tem sido alvo de estudos geoquímicos detalhados, tanto através de interpretação de teores, como mediante a integração de diferentes ferramentas prospectivas de óleos, a exemplo da sísmica 3D e do geoprocessamento de imagens de satélites. Não obstante, a utilização destas ferramentas complementares, o desenvolvimento de técnicas que permitam a caracterização mais imediata e conclusiva de diferentes tipos de óleos, com ênfase nas peculiaridades geoquímicas é um anseio das áreas operacionais que necessitam de ferramentas computacionais para geração de resultados rápidos e precisos.

Há algum tempo a análise visual de cromatogramas tem revelado a existência de similaridades geométricas entre cromatogramas de óleos pertencentes a um mesmo tipo. A detecção de similaridades entre cromatogramas de diferentes campos de uma mesma bacia sedimentar, permitiria avaliar efeitos de compartimentalização de reservatórios, juntamente com informações geológicas da região.

Os perfis dos cromatogramas variam de acordo com a distribuição dos *n*-alcanos nestas amostras. Dentro desse conjunto de amostras existem 6 tipos de cromatogramas de petróleo com perfis bem definidos.

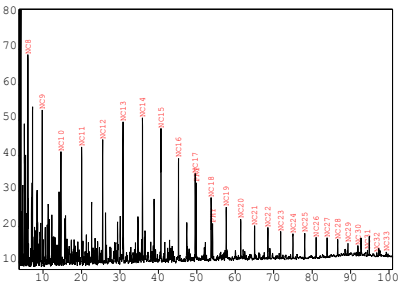
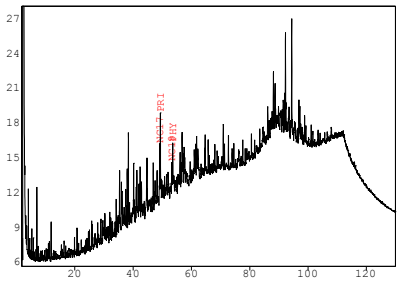
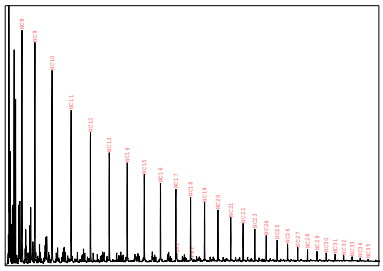
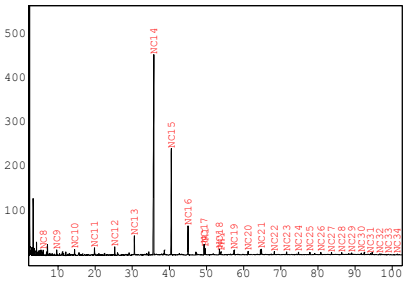
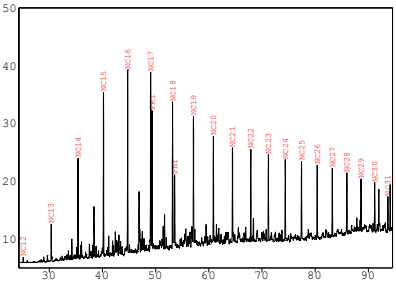
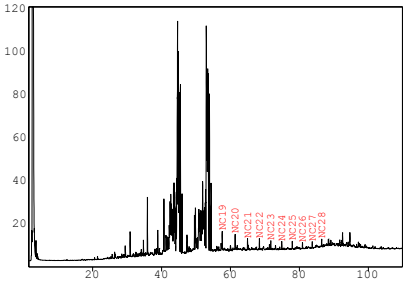
Existem problemas para classificar óleos biodegradados e misturados. Neste caso, outras informações, tais como da linha base e picos intermediários, deveriam ser integrados ao banco de dados, o que o tornaria bem maior. Daí outro interesse em compactar os dados.

A Tabela 1 mostra os tipos de cromatogramas existentes no banco de dados a ser analisado.

2 OBJETIVOS

Este trabalho tem por objetivo aplicar a técnica de classificação por redes neurais e avaliar seu desempenho para um problema de classificação de diferentes tipos de amostras de petróleo, de acordo com os perfis de distribuição dos *n*-alcanos em cromatogramas gasosos e a partir desta rede mensurar seu desempenho quando os dados são compactados. A compactação é feita utilizando-se as técnicas de Análise de Componentes Principais (PCA) e Análise de Relevância das variáveis.

Tabela 1 - Tipos de cromatogramas presentes no banco de dados

Classe	Cromatograma Típico	Classe	Cromatograma Típico
1		4	
2		5	
3		6	

3 DESCRIÇÃO DO MÉTODO DE REDES NEURAIS

Nas redes neurais artificiais, a idéia é realizar o processamento de informações tendo como princípio a organização de neurônios do cérebro. Assim, uma rede neural pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado em aprendizagem (experiência) e disponibilizar este conhecimento para a aplicação em questão.

Existem várias formas de se desenvolver uma rede neural. Ela deve ser montada de acordo com o problema a ser resolvido. Em sua arquitetura são determinados o número de camadas usadas (as camadas são formadas por neurônios), a quantidade de neurônios em cada camada, o tipo de sinapse utilizado, etc.

Resumidamente, uma rede neural consiste em um conjunto de unidades de entrada e saída, onde cada conexão tem um peso associado à ela. A Figura 1 indica a topologia da rede neural utilizada neste trabalho.

Durante a fase de aprendizagem (treinamento), enquanto a rede aprende, ela ajusta os pesos até ser capaz de predizer a correta classe das amostras de entrada. Ou seja, depois da rede treinada, ela é utilizada para fazer predições.

Existem, basicamente, 3 tipos de aprendizado nas redes neurais artificiais. Estaremos trabalhando com o aprendizado supervisionado. Neste tipo, a rede neural recebe um conjunto de entradas padronizados e seus correspondentes padrões de saída, onde ocorrem ajustes nos pesos sinápticos até que o erro entre os padrões de saída gerados pela rede tenham um valor desejado.

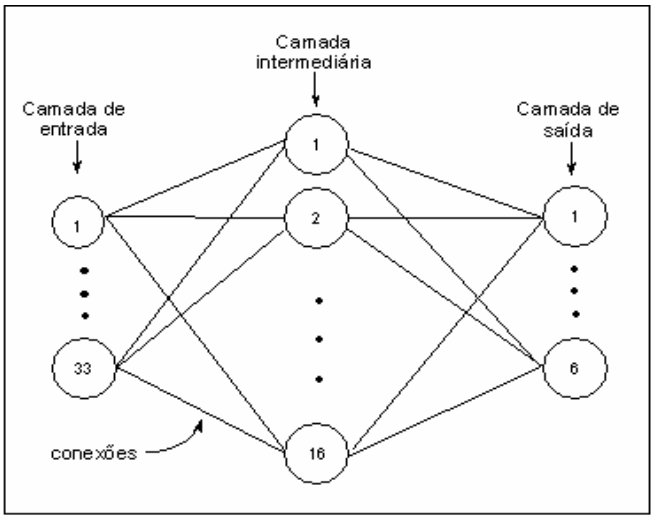


Figura 1 - Configuração da rede neural - tamanho das camadas: entrada - 33, intermediária - 16, saída - 6

4 DESCRIÇÃO DOS MÉTODOS DE COMPACTAÇÃO UTILIZADOS

4.1 Análise de Componentes Principais

A aplicação da Análise de Componentes Principais (PCA) nos problemas é matematicamente simples. Corresponde a calcular autovalores e autovetores do conjunto aleatório de dados e obter as componentes. A PCA é uma transformação linear ortogonal de um espaço p-dimensional para um espaço m-dimensional, m<p. As coordenadas dos dados no novo espaço são não correlacionadas e a maior quantidade de variância dos dados é preservada usando-se apenas algumas poucas coordenadas.

Existem vários métodos de extração de componentes principais. O método aqui utilizado foi o algébrico, que consiste em tornar a matriz de covariância (se a média é nula) ou correlação da base de dados diagonalizável. Para isto calcula-se os autovetores e autovalores da matriz de covariância. Os autovetores representam os vetores unitários na direção de projeção em cada componente e o autovalor está ligado à variância, a energia na direção da componente. De acordo com a energia acumulada na PCA, utiliza-se os dados projetados apenas na direção das componentes principais, as com maior energia, o que fornece uma boa representatividade dos dados já que utiliza-se as componentes mais relevantes.

4.2 Análise de Relevância

Significa "olhar" para as variáveis no meu espaço multidimensional e ver quais são as mais relevantes.

Através da Análise de Relevância das variáveis, com o intuito de se verificar qual a resposta da rede neural quando se elimina variáveis não relevantes, é possível compactar o dado. No entanto, cabe ressaltar que a relevância não consegue verificar correlação entre as variáveis.

O cálculo da relevância aqui utilizado foi efetuado da seguinte maneira (suponha o cálculo da relevância da variável 1):

- calcular a saída da rede para os dados do problema (*out*);
- substituir para todos os eventos o valor da variável 1 pela média da variável 1 e manter os demais;
- calcular a saída da rede para os dados modificados no passo anterior (*out'*);
- calcular o somatório da diferença dos quadrados entre *out* e *out'* (distância euclidiana) para todos os eventos e dividir pelo número de eventos.

Quanto maior for o valor encontrado no cálculo acima (erro), mais relevante é a variável, uma vez que a rede contava com esta variável e quando os valores da variável em todos os eventos foram trocados pela média, a rede forneceu um erro significativo.

5 METODOLOGIA

O banco de dados consiste em uma seqüência dos valores das alturas dos *n*-alcanos homólogos para cada amostra determinados em pico ampére (pA), obtidos como resultados da análise de cromatografia gasosa. O banco de dados original apresenta 33 variáveis como ponto de partida que consistem nos *n*-alcanos dos perfis cromatográficos. A princípio foram utilizadas 300 amostras contendo 6 tipos diferentes de distribuições para serem classificadas por redes neurais. Cada tipo de distribuição foi considerado uma classe, portanto as amostras foram divididas em seis classes, onde 77 amostras eram de Classe 1, 46 da Classe 2, 47 da Classe 3, 64 de Classe 4, 38 de Classe 5 e 28 de Classe 6.

Esses dados foram normalizados, ficando os valores num intervalo entre 0 e 1. Foram selecionadas aleatoriamente cerca de 70% dos dados para treinamento e 30% para teste.

A rede neural foi implementada no *software Matlab*® versão 6.5, utilizando o método do *backpropagation*. Para implementar a rede, diferentes topologias e algoritmos de treinamento foram testados até se chegar ao melhor resultado. Os principais foram: Gradiente descendente; Gradiente descendente com momento; Gradiente descendente com taxa de aprendizado adaptativo; Gradiente descendente com momento e taxa de aprendizado adaptativo; *Resilient Propagation*; Levenberg-Maquardt. Quanto ao critério de parada da rede, na maioria das vezes o treinamento parou devido ao erro no conjunto de validação ter começado a aumentar.

A Análise de Componentes Principais (PCA) e a Análise de Relevância também foram implementadas no *software Matlab*. Para a PCA os dados também foram subtraídos de suas respectivas médias para que ficassem com média nula. Essas duas análises deram origem a outros bancos de dados. São eles: banco original projetado nas componentes principais de acordo com o nível de energia desejado e banco contendo apenas as variáveis relevantes do banco original.

As matrizes de confusão tem por objetivo mostrar a quantidade de amostras classificadas corretamente para cada classe, portanto mede a eficiência. Através dela também pode ser verificado, no caso de percentual de classificação incorreta, com qual classe houve a confusão. Essas matrizes foram calculadas para os conjuntos de treino e teste separadamente e foram implementadas no *Matlab*. Além disso, calculou-se a eficiência média da classificação do conjunto de teste no intuito de se verificar a compactação máxima. A eficiência média é calculada através da soma da diagonal principal da matriz de confusão dividida pelo número de classes.

6 RESULTADOS

Da análise dos diferentes algoritmos utilizados nos bancos de dados, observou-se que aquele que apresentou maior confiabilidade e acurácia foi o algoritmo *Resilient Backpropagation*, utilizando-se 16 neurônios na camada intermediária e funções de ativação do tipo tangente hiperbólica. Neste caso, a convergência se deu em média por volta de 50 épocas.

Os resultados encontrados para a Análise de Componentes Principais podem ser vistos na Tabela 2 e nas Figuras 2 e 3.

Tabela 2 - Energia Acumulada

Nº de componentes	Energia Acumulada (em %)
1	74,993
2	84,352
3	90,696
4	95,067
5	96,721
6	97,73
7	98,351
8	98,65
9	98,917
10	99,157
⋮	⋮

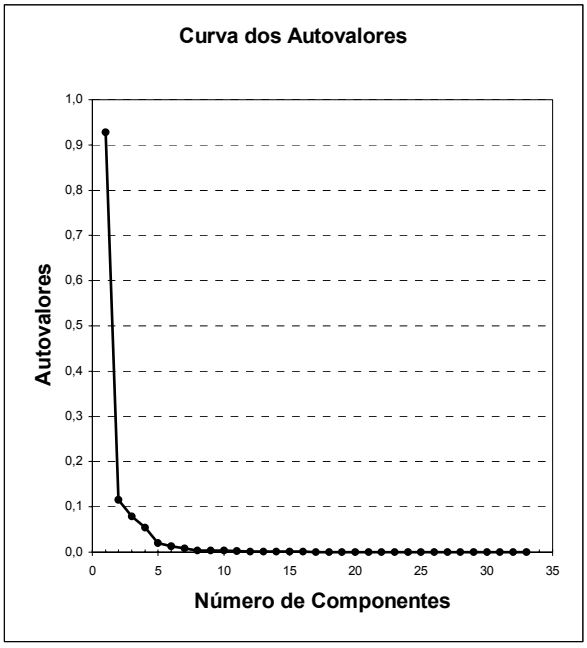


Figura 2 - Curva dos autovalores

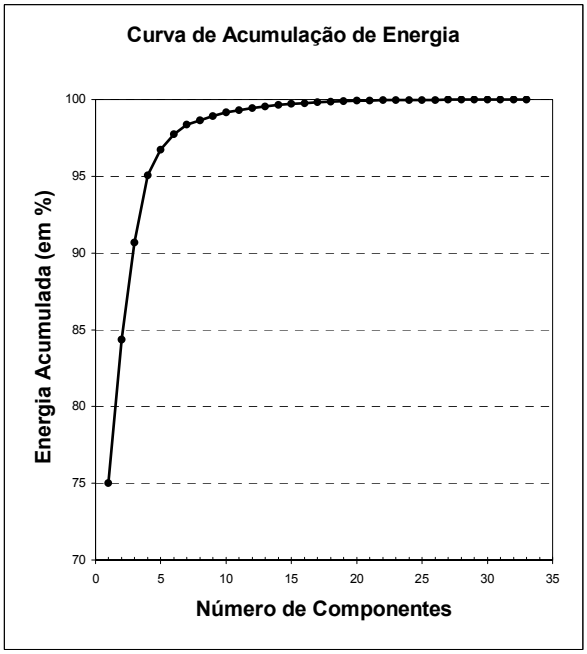


Figura 3 - Curva de acumulação de energia

Na Tabela 3 pode-se verificar os valores encontrados para as relevâncias das variáveis. Baseado nos valores encontrados, efetuou-se uma padronização para que a relevância ficasse evidente. Considerando-se como variáveis não relevantes as com valores menores que 0.1, 10 variáveis puderam ser removidas, restando 23 variáveis. Criou-se um novo banco de dados apenas com estas variáveis relevantes e o algoritmo de classificação foi aplicado a este novo banco, onde se pôde verificar que a eficiência média no conjunto de teste foi de 94,97% e as eficiências de todas as classes também foram excelentes.

Tabela 3 - Relevância obtida no *Matlab*

Variáveis	Relevância	Relevância Padronizada
Var_1	1,7629	0,79388
Var_2	1,637	0,73719
Var_3	1,6107	0,72534
Var_4	0,79846	0,35957
Var_5	0,20097	0,09050
Var 6	0,1548	0,06971
Var7	2,2206	1,00000
Var 8	0,56094	0,25261
Var 9	0,72547	0,32670
Var 10	0,7882	0,35495
Var 11	0,80397	0,36205
Var 12	0,6651	0,29951
Var 13	0,80363	0,36190
Var 14	0,034813	0,01568
Var 15	0,1728	0,07782
Var 16	0,0057674	0,00260
Var17	0,68759	0,30964
Var 18	0,63679	0,28676
Var 19	0,78523	0,35361
Var 20	0,83185	0,37461
Var 21	0,7139	0,32149
Var 22	0,77387	0,34850
Var 23	0,75424	0,33966
Var 24	0,378	0,17022
Var 25	0,46399	0,20895
Var 26	0,4625	0,20828
Var_27	0,47196	0,21254
Var_28	0,028931	0,01303
Var_29	0,029718	0,01338
Var_30	0,010436	0,00470
Var_31	0,1135	0,05111
Var_32	0,0014493	0,00065
Var_33	0,40302	0,18149

As matrizes de confusão dos conjuntos de treino (MCTR) e de teste (MCTS) calculadas para o banco de dados original, banco de dados projetado em 10 componentes e banco de dados pós-análise de relevância (onde as variáveis não relevantes foram eliminadas como descrito anteriormente), podem ser vistas nas Tabelas 4, 5 e 6, respectivamente.

Tabela 4 - MCTR e MCTS da base de dados original

MCTR	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**100**	0	0	0	0	0
Classe 2	10,256	**89,744**	0	0	0	0
Classe 3	0	0	**100**	0	0	0
Classe 4	0	2,2727	**0**	**95,455**	2,2727	0
Classe 5	0	4	**0**	0	**96**	0
Classe 6	0	0	**12,5**	0	0	**87,5**

MCTS	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**95,652**	4,3478	0	0	0	0
Classe 2	14,286	**85,714**	0	0	0	0
Classe 3	0	0	**100**	0	0	0
Classe 4	0	0	0	**95**	5	0
Classe 5	0	0	0	0	**100**	0
Classe 6	0	0	16,667	0	0	**83,333**

Tabela 5 - MCTR e MCTS da base de dados projetada em 10 componentes

MCTR	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**100**	0	0	0	0	0
Classe 2	7,547	**92,453**	0	0	0	0
Classe 3	0	7,5472	**92,453**	0	0	0
Classe 4	0	0	0	**98,113**	1,8868	0
Classe 5	0	0	0	0	**100**	0
Classe 6	0	0	0	0	0	**100**

MCTS	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**100**	0	0	0	0	0
Classe 2	14,286	**85,714**	0	0	0	0
Classe 3	0	0	**100**	0	0	0
Classe 4	0	0	0	**94,118**	5,882	0
Classe 5	0	0	0	0	**90**	10
Classe 6	0	0	0	0	0	**100**

Tabela 6 - MCTR e MCTS da base de dados pós-análise de relevância

MCTR	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**100**	0	0	0	0	0
Classe 2	7,547	**92,453**	0	0	0	0
Classe 3	0	0	**100**	0	0	0
Classe 4	0	0	0	**98,113**	1,8368	0
Classe 5	0	0	0	0	**100**	0
Classe 6	0	0	0	0	0	**100**

MCTS	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Classe 1	**100**	0	0	0	0	0
Classe 2	14,286	**85,714**	0	0	0	0
Classe 3	0	0	**100**	0	0	0
Classe 4	0	0	0	**94,118**	5,8324	0
Classe 5	0	0	0	10	**90**	0
Classe 6	0	0	0	0	0	**100**

Obteve-se para eficiência média no treinamento um percentual de 94,78 %, 97,17 % e 98,43 % e no teste 93,28 %, 94,972 % e 94,972 % para a base de dados original, base projetada em 10 componentes e base pós-análise de relevância, respectivamente. Os erros de classificação estão dentro dos limites aceitáveis (abaixo de 10%).

Pode-se observar na Figura 4 qual a compactação máxima, quando os dados foram projetados em 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30 e 33 componentes. Observa-se que a partir de 10 componentes a eficiência começa a cair. Esta análise indica ser interessante não utilizar menos de 10 componentes, já que para uma quantidade menor, embora a eficiência média tenha sido satisfatória, quando observa-se a eficiência em cada classe, verifica-se um aumento no erro. Além disso, para 10 componentes a eficiência foi máxima em todas as classes. Logo, a compactação máxima consiste em utilizar 10 componentes.

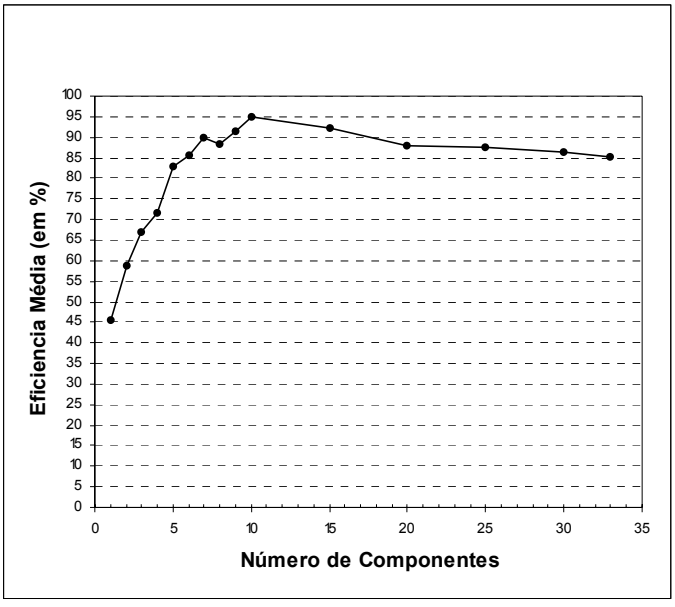


Figura 4 - Curva de eficiência média do conjunto de teste em função do número de componentes

Cabe ressaltar que uma análise preliminar já demonstrou que uma outra técnica de classificação, como a árvore de decisão, mostrou uma acurácia menor do que a obtida utilizando-se redes neurais. Sendo assim, deve ser feita uma análise mais aprofundada dos métodos para confirmar esse resultado.

7 CONCLUSÕES

A Classificação neural de diferentes tipos de amostras de óleo baseada em cromatogramas gasosos apresentou um resultado excelente. Neste caso, o algoritmo mais eficiente foi o *resilient backpropagation*.

Embora a utilização de três componentes principais já represente 90,696% da variabilidade dos dados, quando foi feita uma análise mais detalhada considerando-se as eficiências médias nos conjuntos de teste em função das componentes principais, como pöde-se verificar na Figura 4, a compactação máxima de modo que a eficiência mantenha-se boa consiste em utilizar 10 componentes.

Após a análise de relevância efetuada no *Matla*b, pôde-se verificar que a eficiência média no conjunto de teste foi de 94,97% e as eficiências de todas as classes também foram excelentes. O mesmo aconteceu quando os dados foram projetados em 10 componentes.

As duas técnicas de compactação utilizadas neste trabalho foram bastante eficientes. Com a integração de novas informações ao banco de dados essas técnicas serão muito úteis.

Além disso, aprimoramentos na configuração da rede neural utilizada, assim como a aplicação de outras técnicas de classificação se fazem pertinentes, com a finalidade do aumento da acurácia na classificação, uma vez que é imprescindível uma caracterização precisa do sistema petrolífero durante a exploração.

8 AGRADECIMENTOS

Ao Laboratório de Modelagem de Bacias Sedimentares (LAB2M), Ao Laboratório de Métodos Computacionais em Engenharia (LAMCE), a Agência Nacional do Petróleo (ANP) e a Petrobrás (Cenpes) pelo suporte técnico e financeiro.

9 REFERÊNCIA BIBLIOGRÁFICA

Haykin, S., 2001, Redes Neurais: Princípios e Prática, 2ª edição, Bookman.

Fausset, L., 1984, Fundamentals of Neural Networks, Prentice Hall.

Kovács, Z., 1996, Redes Neurais Artificiais, Acadêmica.

Serra, A. C. S. 2003. A influência de aditivos de lamas de perfuração sobre as propriedades geoquímicas de óleos. Tese de M. Sc., COPPE/UFRJ, Rio de Janeiro, Brasil.

Collins, C.H., Braga, G.L., Bonato, P.S., 1997, Introdução a Métodos Cromatográficos, Editora Unicamp, São Paulo.

Aquino Neto, F.R., Nunes, D.S.S., 2003, Cromatografia: Princípios básicos e técnicas afins, Editora Interciência, Rio de Janeiro.