Uma nova métrica para redução de dimensionalidade em modelos de aprendizado neural

Sandro da Silva Camargo[1], Paulo Martins Engel[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{scamargo, engel}@inf.ufrgs.br

Abstract. Cientistas das diversas áreas de conhecimento humano têm se deparado com a necessidade de resolver problemas que envolvem a compreensão de dados de alta dimensionalidade. Assim, a utilização de técnicas de aprendizado de máquina tem sido dificultada por limitações na quantidade de exemplos aliada ao excesso de variáveis aleatórias descrevendo os dados. Este trabalho apresenta uma nova abordagem para redução de dimensionalidade dos dados a fim de melhorar a escalabilidade dos algoritmos de aprendizado baseados em redes neurais. São apresentados experimentos que demonstram a eficiência da abordagem proposta na criação de modelos neurais mais simples e mais precisos dos dados.

Keywords: Redução de dimensionalidade, aprendizado de máquina, redes neurais, modelos neurais.

1 Introdução

O atual estágio de evolução científica tem conduzido ao constante desenvolvimento de novas tecnologias de geração e aquisição de dados e também à proliferação de recursos para obtenção de dados através de simulação computacional. Concomitantemente, o custo de armazenamento de dados tem caído continuamente. Tais fatores contribuem para que a quantidade de informação armazenada dobre aproximadamente a cada 20 meses [8]. Como conseqüência desta realidade, observa-se um crescimento exponencial dos bancos de dados das diversas áreas de conhecimento humano, sejam elas: científicas, comerciais ou governamentais. Na área científica, além da grande quantidade observações que constituem a população sob análise, também tem se tornado comum a alta dimensionalidade, ou seja, a descrição de cada uma das observações em relação a um grande número de variáveis aleatórias. O senso comum sugere que quanto maior a quantidade de dados obtida sobre um determinado fenômeno ou evento, maior será a compreensão sobre ele e, consequentemente, maior a capacidade de predizer seu comportamento futuro. Porém, sob um ponto de vista estatístico, o excesso de variáveis aleatórias descrevendo cada observação pode geralmente conduz a uma maior dificuldade de compreensão dos padrões expressos na população.

O problema da alta dimensionalidade foi abordado pela primeira vez, em [2], que definiu empiricamente que bancos de dados nos quais a proporção de variáveis/observações for maior que 0,1, podem ser considerados de alta dimensionalidade. Desde então, diversas abordagens foram propostas na literatura para diminuir os efeitos da alta dimensionalidade. Com base na estratégia que utilizam, estas abordagens podem ser agrupadas em: abordagens de seleção, extração ou construção de variáveis [5]. No tocante a seleção de variáveis, diversas métricas de similaridade distintas são utilizadas para identificar as mais relevantes, sendo as mais populares apresentadas em [12].

Neste trabalho é proposta uma nova abordagem para redução de dimensionalidade dos dados (RDD) que utiliza a abordagem de seleção de variáveis. Esta abordagem é baseada em uma heurística que, através de um modelo baseado em redes neurais, iterativamente refina o modelo preditivo. A abordagem proposta é aplicada a dois problemas distintos, sendo um sintético e outro real, a fim de demonstrar sua superioridade em relação a outras abordagens similares propostas na literatura. Este trabalho está organizado da seguinte forma: na seção 2 são apresentados alguns conceitos básicos sobre RDD e o problema é formalizado. Posteriormente é apresentada a abordagem proposta, que é aplicada a um problema clássico de aprendizado, e seus resultados comparados com as métricas de similaridade mais populares. Finalmente, na seção 4 são apresentados os resultados obtidos através da aplicação da abordagem proposta em um problema real e demonstrado o ganho de exatidão por ela propiciado.

2 O problema da dimensionalidade

Em dados de alta dimensionalidade, a criação de modelos preditivos dos dados necessita considerar espaços de busca inerentemente esparsos [5]. Para vencer esta dificuldade, os cientistas constantemente têm se deparado com a necessidade de encontrar estruturas significativas ocultas, de baixa dimensão, dentro de observações de alta dimensão, sendo tal técnica de pré-processamento denominada de redução de dimensionalidade dos dados. Analogamente, o cérebro humano se confronta com o mesmo problema em suas percepções diárias, extraindo, de forma eficiente, um pequeno número de estímulos relevantes a partir de aproximadamente 30.000 fibras nervosas sensoriais [10].

A RDD pode ser dividida em três categorias independentes e igualmente difíceis: seleção de variáveis, extração de variáveis e construção de variáveis. A seleção de variáveis, ou seleção de subconjunto de atributos (SSA), é especialmente útil quando há uma grande quantidade de variáveis caracterizando cada observação no banco de dados, propriedade peculiar a diversos bancos de dados, notoriamente os científicos. Nestes casos, a quantidade de observações necessárias para adaptar um modelo multivariado cresce exponencialmente em relação à quantidade de variáveis que representam cada observação. Além disso, o uso de muitas variáveis no modelo preditivo pode dificultar a interpretação da análise e viola o princípio da parcimônia. Outro fator importante é que a muitas variáveis podem mais facilmente conduzir a uma superadaptação do modelo preditivo[5]. Embora os algoritmos de mineração de

dados já apliquem internamente a seleção das variáveis mais informativos, ignorando os menos informativos, a utilização de algoritmos de SSA geralmente melhora o desempenho destes algoritmos [11]. Segundo [13], os objetivos da SSA em aprendizado de máquina são: 1) Reduzir a dimensionalidade do espaço de variáveis; 2) Acelerar o aprendizado dos algoritmos de mineração de dados; 3) Melhorar a capacidade preditiva dos algoritmos; e 4) Melhorar a compreensibilidade dos resultados obtidos.

Formalmente, o problema de SSA pode ser definido como o processo de encontrar um conjunto relevante de M variáveis dentre as N variáveis originais, onde $M <= N$, para definir os dados a fim de maximizar a exatidão preditiva do modelo [7]. A eficiência das abordagens de RDD é altamente dependente da natureza do problema que se busca aprender. Algumas abordagens tais como entropia ou coeficiente de correlação de Pearson, podem ser extremamente eficientes em problemas de natureza linear e pouco eficientes em problemas de natureza não linear. Já as RNAs tem a capacidade de aprenderem problemas tanto lineares quanto não lineares.

3 Abordagem Proposta

Neste trabalho é proposta uma abordagem de RDD para problemas de aprendizado supervisionado, utilizando redes neurais, baseado na ordenação da importância dos variáveis de entrada. Esta proposta visa elucidar a seguinte hipótese de pesquisa: "Os pesos das sinapses que ligam a camada de entrada à primeira camada oculta podem ser utilizados para reduzir a dimensionalidade dos dados de entrada para o aprendizado neural." Baseado nesta hipótese, propõe-se que a definição da importância de cada atributo seja dada por um escore que se baseia nos pesos sinápticos da camada de entrada da rede. A partir da definição da importância que cada atributo de entrada tem na predição do valor da saída da rede neural é então proposta uma abordagem de RDD para otimizar a criação dos modelos neurais.

Segundo [1], em uma rede neural multicamada, as sinapses que ligam a camada de entrada à primeira camada oculta são responsáveis por codificarem os padrões expressos nos dados. Já as sinapses que ligam as camadas posteriores são responsáveis por decodificarem os padrões, a fim de gerar uma saída para a rede. Baseado nesta idéia, decidiu-se que a definição do escore proposto somente consideraria os pesos sinápticos entre a camada de entrada e a primeira camada oculta.

Desta forma, o cálculo do escore seria dado pela seguinte fórmula:

$$s_i = \frac{\sum_{k=1}^{n} | w_{ik} |}{x} \qquad (1)$$

Tendo-se que:

- O escore da variável aleatória i é dado por s_i ;

- Há x neurônios na primeira camada oculta;
- w_{ik} é o peso da sinapse entre o i-ésimo neurônio da camada de entrada e o k-ésimo neurônio da primeira camada oculta;

A abordagem proposta é composta por 6 passos. No passo 1 é realizado o treinamento convencional da rede neural. No passo 2 são calculados os escores a partir da fórmula proposta. No passo 3 é realizada a avaliação dos subconjuntos de atributos através da criação de modelos incrementais, a partir do modelo mais simples com um atributo, de forma que cada novo modelo contenha um atributo a mais que o modelo anterior. A cada novo atributo acrescentado ao modelo, é realizada a RDD de treinamento, a rede neural é treinada, e seu resultado é avaliado. Este processo é realizado iterativamente até que seja atingido o critério de parada. No passo 4, os dados de entrada são reduzidos através da eliminação das características que não pertencem ao melhor subconjunto. No passo 5 a rede neural é treinada novamente tal como foi realizado no passo 1. E finalmente, no passo 6, o modelo é avaliado segundo as métricas convencionais a fim de avaliar a qualidade do modelo com dimensionalidade reduzida.

Tendo-se que os dados originais são representados em uma matriz t x a, onde t é a quantidade de observações e a é a quantidade de atributos descrevendo cada observação, a abordagem proposta visa transformar os dados em uma nova matriz t x b, onde $b < a$. Os b atributos selecionados para este modelo são aqueles com maiores escores, e a intenção é que o modelo com b atributos de entrada atinja um maior nível de exatidão preditiva que o modelo com a atributos.

3.1 Evidências Experimentais

A fim de se obter evidências experimentais do sucesso da abordagem proposta, foi gerado um banco de dados que expressa a função não-linear do 'ou exclusivo' (XOR). Esta função foi escolhida por representar um padrão não linearmente separável, o que também é um aspecto peculiar a vários problemas reais [4]. A fim de identificar se a abordagem proposta é capaz de reduzir a dimensionalidade dos dados de entrada, foram incluídas como entradas da rede neural, além das duas variáveis usadas como entrada para a função XOR, mais diversas variáveis com ruído.

3.1.1 Banco de Dados

O banco de dados gerado é composto por 20 variáveis e 100 observações. Desta forma, a dimensionalidade, ou proporção variáveis/observação, deste banco é 0,2. Cada variável, para uma dada observação, pode assumir aleatoriamente os valores 0 ou 1. O conjunto de entradas *input_data* é dado por:

$$\text{input\_data} = \text{round}(\text{rand}(\text{features},\text{samples})) \quad (2)$$

onde *features* é o número de variáveis aleatórias para cada observação, definido como 20; e *samples* é o número de observações do conjunto de dados, definido como 100. Foram gerados diversos conjuntos de dados com diferentes valores de *features* e

samples. Nos experimentos aqui relatados foram definidos os valores 20 e 100 pois permitiam um bom balanceamento entre quantidade de observações e dificuldade de aprendizado do problema. Uma maior quantidade de observações tornaria o problema proporcionalmente mais fácil de ser aprendido pela rede. Uma menor quantidade de observações faria com que a rede não conseguisse aprender absolutamente nada sobre o problema, sendo gerado um modelo pior que o modelo de aprendizado fraco [4].

Os valores de saída para este conjunto de observações são dados por uma função XOR dos valores da primeira e da segunda coluna:

$$output\_data = \text{xor}(input\_data(1,:), input\_data(2,:)) \quad (3)$$

Adicionalmente, tanto os valores de entrada quando os de saída são escolanados de forma que o valor mínimo seja -1 e o valor máximo seja 1. A função de escalonamento é dada pela seguinte fórmula:

$$pe = 2*(p-minp)/(maxp-minp) - 1 \quad (4)$$

onde *minp* é o valor mínimo assumido pela variável em todas as observações, *maxp* é o valor máximo assumido pela variável em todas as observações, *p* é o valor do atributo na observação atual, e *pe* é o valor *p* escalonado.

3.1.2 Resultados

Para modelar este problema foi utilizada uma RNA treinada com o algoritmo Resilient Backpropagation proposto por [9]. Também foram realizados experimentos com outros algoritmos de treinamento, sendo obtidos resultados similares. Para obter-se uma significância estatística, os experimentos foram repetidos 20 vezes sobre condições idênticas de execução, exceto pelo reinício aleatório dos pesos sinápticos. O processo de validação cruzada foi realizado utilizada a técnica 10-fold.

Tabela 1. Distância euclidiana entre os valores ideais de escore e os valores obtidos com cada uma das abordagens utilizadas.

Abordagem	Distância
Escore proposto	0.28812
Mahalanobis	1.69662
Generalized Least Squares	2.12132
Ordinary Least Squares	2.12132
Internal Product	2.12132
Covariance	2.12132
Kendall	2.12314
Spearman	2.12314
Correlation coefficient	2.12314
T Test Regression	2.80792
Regression	2.80792
Welch Test	2.88818

T Test	2.88818
Wilcoxon	2.94615
U Test	2.95901
Kruskal Wallis	2.95901
Sign	2.99669
Chi-square	3.01954
Var Test	3.70111
Bartlett	4.01253

Adicionalmente, foram calculados os escores de cada um dos 20 atributos pela abordagem proposta e por outras abordagens descritas na literatura. Dado que somente 2 dos 20 atributos são entradas relevantes para a função XOR, os escores ideais seriam: escore máximo "1" para os dois atributos relevantes; e escore mínimo "0" para os demais atributos. Os escores finais dos atributos, após aplicação das diversas abordagens, foram escalonados no intervalo [0,1]. A eficiência de cada uma das abordagens foi baseada na distância euclidiana entre os escores obtidos e os escores ideais. A tabela 1 mostra que a abordagem conseguiu chegar mais próxima dos escores ideais que as demais abordagens utilizadas.

Calculados os escores dos atributos através da abordagem proposta, foi realizada a RDD sobre o banco de dados para verificar a eficiência de todo o processo e o ganho de precisão obtido. Enquanto o modelo neural criado com todos os 20 atributos de entrada obteve 28% de taxa de erro, o modelo criado após a aplicação da abordagem proposta obteve 2,85% de taxa de erro, conseguindo ser mais exato com uma menor quantidade de atributos. Os relatórios gerados durante os processos de treinamento, com e sem RDD, estão disponíveis respectivamente em: www.inf.ufrgs.br/~scamargo/DR.

Também foram realizados experimentos em um problema linear onde a abordagem proposta mostra-se pouco superior as outras abordagens aplicadas. Os resultados encontram-se disponíveis no endereço acima mencionado.

4 Avaliação de Desempenho

A avaliação de desempenho da abordagem proposta foi realizada a partir de resultados obtidos em experimentos de regressão e classificação sobre conjuntos de dados reais. Uma forma sistemática de avaliação de desempenho é importante para que sejam obtidos resultados confiáveis permitindo a comparação e, principalmente, a replicação dos experimentos realizados.

4.1 Regressão em um conjunto de dados real

Nestes experimentos buscou-se desenvolver modelos de qualidade dos reservatórios os dados e interpretações produzidas no estudo dos arenitos da Formação de Uerê, Devoniano da Bacia do Solimões [6]. Este estudo sistemático e detalhado definiu, através da combinação de observações petrográficas e dados petrofísicos de porosidade e permeabilidade, petrofácies de reservatório com significância coerente e geneticamente suportada dos reservatórios Uerê. O conjunto de dados e interpretações gerados naquele estudo irá, portanto, constituir uma base ideal para a definição de modelos de qualidade de reservatórios através de técnicas de descoberta de conhecimento, bem como para o teste da precisão e eficiência dessas técnicas para o desenvolvimento sistemático de modelos operacionais de caracterização e previsão de qualidade de reservatórios clásticos de hidrocarbonetos. A abordagem proposta foi aplicada em [3] e os resultados obtidos são descritos a seguir.

4.1.1 Dados

Os dados utilizados, que foram publicados por [6], contém dados e interpretações sobre arenitos Devonianos da formação de Uerê, que são importantes alvos de exploração de petróleo na Bacia do Solimões, na Amazônia Brasileira.

Estes dados são compostos por 58 observações, cada uma contendo 92 atributos descrevendo parâmetros petrográficos e petrofísicos. Dos 92 atributos de entrada, algumas foram excluídas por orientação do especialista do domínio por conterem resultados de análises posteriores, que tornariam trivial a tarefa de predição. Durante o pré-processamento foram eliminados 32 atributos, ficando os dados finais com 60 atributos. Assim, a dimensionalidade dos dados é 1,034.

4.1.2 Resultados

Para modelar este problema foram utilizadas RNAs treinada com o algoritmo Resilient Backpropagation proposto por [9]. Para obter-se uma significância estatística, os experimentos foram repetidos 20 vezes sobre condições idênticas de execução. O processo de validação cruzada foi realizado utilizada a técnica leave-one-out, devido à pequena quantidade de observações. A figura 1 mostra os escores obtidos para cada um dos 60 atributos de entrada.

A partir da identificação dos atributos com maior escore foram então gerados e avaliados os 60 subconjuntos de atributos, sendo então mostrados na tabela 2 alguns dos resultados mais relevantes. Na tabela estão grifados o modelo com todos os atributos e o melhor modelo obtido com a aplicação da abordagem proposta, que continha 3 atributos de entrada.

Tabela 2. Variação da taxa de erro e da dimensionalidade em função do número de atributos de entrada.

Atributos de Entrada	Dimensionalidade	Erro Absoluto	Erro Percentual

1	0,017	2,7173	26,17%
2	0,034	2,7110	26,11%
3	0,051	1,8637	17,97%
4	0,068	1,9578	18,86%
5	0,086	2,0533	19,78%
60	1,034	2.1403	20,61%

A figura 2 apresenta os erros obtidos com o melhor modelo gerado através da aplicação da abordagem proposta, e o modelo com todos os atributos de entrada.

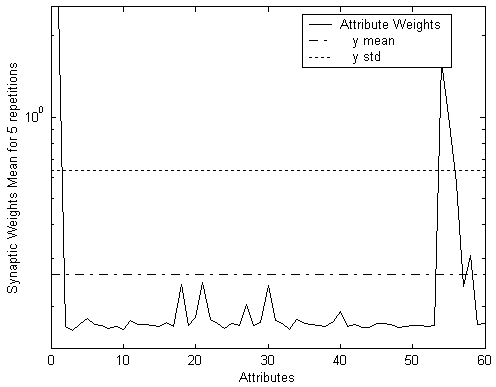

Fig. 1. Escores para cada um dos 60 atributos de entrada.

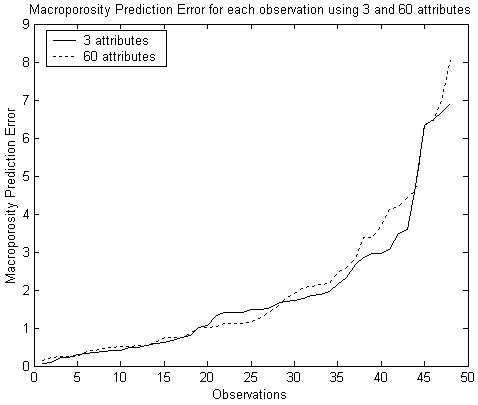

Fig. 2. Erros de predição de macroporosidade com 3 e 60 atributos de entrada.

Adicionalmente, as figuras 3 e 4 apresentam os resultados dos modelos com 3 e 60 atributos, mostrando que o modelo gerado com a aplicação da abordagem proposta consegue obter um melhor coeficiente de correlação com os valores reais, do que o modelo original com todos os atributos de entrada. Enquanto o modelo utilizando

todas as variáveis de entrada permite à rede neural atingir um coeficiente de correlação de 0,8797 com os valores reais, o modelo reduzido criado com a abordagem proposta obtém resultados mais precisos, atingindo um coeficiente de correlação de 0,8927.

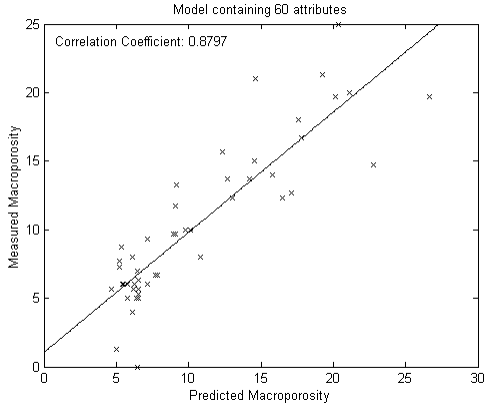

Fig. 3. Correlação entre os valores preditos com o modelo de 60 atributos e os valores reais.

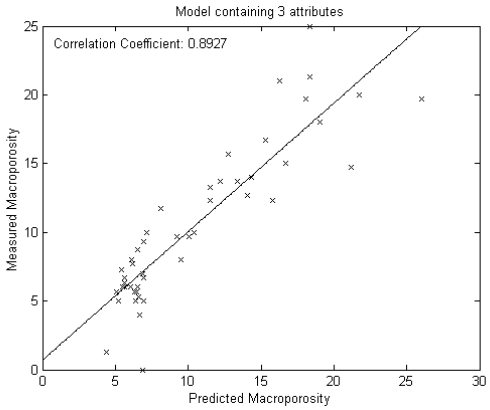

Fig. 4. Correlação entre os valores preditos com o modelo de 3 atributos e os valores reais.

5 Conclusões

A abordagem proposta foi desenvolvida dentro do escopo do projeto DC3PA [3], a fim de superar as dificuldades causadas pela alta dimensionalidade dos dados sendo tratados. Dentro deste escopo de problema, a abordagem conduziu a uma RDD do problema original, permitindo reduzir a dimensionalidade dos dados de entrada, e conduzindo à criação de modelos com menor quantidade de variáveis e com uma maior exatidão preditiva. As maiores contribuições da abordagem proposta residem

nos seguintes aspectos: 1) É intuitiva e de fácil aplicação, 2) Pode ser aplicada de maneira idêntica tanto em problemas de regressão quanto de classificação, 3) Pode ser integrada ao processo de aprendizado neural de maneira transparente e sem a necessidade de configuração de parâmetros adicionais. 4) Conforme demonstrado na subseção 3.1.2, é eficiente em problemas de natureza não linear. As contribuições 2, 3 e 4, consistem em restrições proibitivas para aplicação de muitas outras abordagens similares propostas na literatura.

Adicionalmente, a abordagem também foi aplicada a alguns problemas de predição de séries temporais, sendo bem sucedida.

Acknowledgments. Este trabalho está dentro do escopo do projeto Descoberta de Conhecimento sobre Parâmetros Petrográficos e Petrofísicos de Arenito reservatórios (DC3PA), financiado pelo MCT/CNPq.

References

1. Alpaydin, E. Introduction to Machine Learning. MIT Press, Cambridge (2004)
2. Bellman, R. Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
3. Engel, P. M. Criação de Modelos da Qualidade de Reservatórios pela Aplicação de Técnicas de Descoberta de Conhecimento sobre Parâmetros Petrográficos e Petrofísicos de Arenitos – DC3PA. Projeto de Pesquisa – Instituto de Informática, UFRGS, Porto Alegre (2005)
4. Haykin, S. Neural networks: a comprehensive foundation. Prentice-Hall, Delhi (1999)
5. Larose, D. T. Data mining methods and models. John Wiley & Sons, New Jersey (2006)
6. Lima, R.L., De Ros, L.F. "The role of depositional setting and diagenesis on the reservoir quality of Devonian sandstones from the Solimões Basin, Brazilian Amazonia". In: Marine and petroleum geology, 19, 9, 1047--1071 (2002)
7. Liu, H., Setiono, R. Feature Selection and Classification – A Probabilistic Wrapper Approach. In Proceedings of the Ninth International Conference on Industrial and Engineering Applications of AI and ES, Fukuoka, Japan. 419--424 (1996)
8. Maimon, O. e Rokach, L. Soft computing for knowledge discovery and data mining. Springer, New York (2008)
9. Riedmiller, M., Braun, H. A direct adaptive method for faster backpropagation learning: The Rprop algorithm. Proceedings of the IEEE International Conference on Neural Networks, IEEE Press, 586--591 (1993)
10. Tenenbaum, J. B., De Silva, V., e Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. In Science Magazine, 290, 5500, 2319--2323 (2000)
11. Witten, A. A., Frank, E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco (2005)
12. Yampolskiy, R. V., Govindaraju, V. Similarity Measure Functions for Strategy-Based Biometrics. Proceedings of World Academy of Science, Engineering and Technology, 18, 174--179 (2006)
13. Ye, N. Handbook of Data Mining. Lawrence Erlbaum Associates Publishers, London (2003)