Calibração de modelos ambientais a partir de meta-modelos baseados em classificadores

Flávio C. Oliveira[1], Britaldo S. Soares-Filho[2], Sérgio D Faria[2]

[1] Programa de mestrado em Análise e Modelagem de Sistemas Ambientais – Universidade Federal de Minas Gerais (UFMG)
Av. Anônio Carlos 6627 – 31.270-930 – Belo Horizonte – MG – Brazil

[2]Departamento de Cartografia (IGC) – Universidade Federal de Minas Gerais (UFMG)
Av. Anônio Carlos 6627 – 31.270-930 – Belo Horizonte – MG – Brazil

foliverbr@gmail.com, britaldo@csr.ufmg.br, sergiofaria@ufmg.br

Abstract. *A well performed numerical environmental model calibration process is primordial in achieving a more reliable system. Due to the inherent time-consuming problem, many uses Genetic Algorithms (GAs) to speed up the calibration. Such heuristic is considered well suited to tackle the problem, however, for some models, evaluating the fitness function means running the model thousands of times leading to an impracticable situation. This paper describes how GAs can make use of classifiers K-NN, MSVM among others to build meta-models avoiding the overall number of model runs. The use of classifiers shows that it is possible to significantly reduce the actual fitness evaluation without losing solution quality.*

Resumo. *O cuidado em se realizar uma boa calibração de modelos ambientais numéricos é fundamental na obtenção de sistemas que sejam confiáveis. Devido ao inerente problema de tempo gasto no processo, algoritmos genéticos (AGs) são usados na tentativa de acelerar a calibração. Tal heurística é considerada como sendo apropriada para se tratar o problema, porém, para alguns modelos, avaliar a função fitness significa executar o modelo milhares de vezes, o que pode ser infactível. Este trabalho descreve como AGs podem fazer uso de classificadores K-NN, MSVM entre outros na construção de meta-modelos evitando portanto o número total de avaliações do modelo. O uso de classificadores mostra que é possível reduzir significantemente tais execuções sem que a qualidade da solução seja perdida.*

1. Introdução

A questão ambiental constitui-se hoje em um dos maiores desafios com que a humanidade se depara, neste sentido, a modelagem ambiental é ferramenta fundamental na exploração desses sitemas, possibilitando não somente um melhor entendimento dos processos físicos envolvidos como também permitindo a previsão de eventos e, como consequência, um melhor gerenciamento dos recursos naturais e processos ambientais. Porém, o estudo de modelagem de sistemas reais tem se mostrado uma área complexa e cheia de desafios tanto no que diz respeito à aproximação dos modelos da realidade quanto na complexidade computacional que esses modelos apresentam. O processo de modelagem de sistemas ambientais pode ser dividido em diversas etapas, uma delas é a de calibração do modelo. Segundo H. Madsen (2000) esta etapa consiste na alterações nos valores dos parâmetros de entrada do modelo definidos na etapa de parametrização

no intuito de que a saída do modelo represente bem um certo conjunto de dados observados e aumentem a confiança do modelo. Porém, a quantidade de parâmetros a serem ajustados pode inviabilizar a calibração já que o tempo gasto na verificação de todas as possibilidades de valores para os parâmetros é na maioria das vezes infactível. Sendo assim, várias técnicas existem para tornar a etapa de calibração menos custosa computacionalmente ao mesmo tempo que permitem obter soluções de boa qualidade. H. Madsen (2000), P.O., Yapo et al (1989) e Yang Liu (2005) observam que algoritmos genéticos são bem sucedidos nesta tarefa, por se tratar de um método de otimização genérico (uma meta-heurística) que é capaz de tratar vários problemas diferentes independente da função objetivo a ser otimizada. Os algoritmos genéticos fazem parte de uma classe de algoritmos denominados algoritmos evolucionários e portanto se baseiam nos conceitos de evolução da biologia. Assim sendo, o algoritmo tem como saída indivíduos que são na verdade soluções para o problema que se deseja otimizar e a idéia é que a partir do cruzamento de indivíduos, novos indivíduos melhores surjam e assim uma solução satisfatória do problema seja encontrada. Porém, verifica-se que a avaliação da qualidade dos indivíduos, que é feito avaliando-se a função de adaptação, muitas vezes requer a execução de todo o modelo numérico milhares de vezes, Yaochu Jin (2005) cita que modelos de dinâmica de fluidos computacional CFD (*Computational Fluid Dynamics*, da sigla em inglês) para o caso 3D podem demorar mais de 10 horas em computadores de alta eficiência para executar apenas uma vez, tornando o processo de calibração por métodos tradicionais como GAs e MOGAs inviáveis. Por esse motivo, Yang Liu e S.T. Khu (2007) destacam que pesquisadores já reconhecem a necessidade de uma abordagem eficiente ao problema de tempo gasto na utilização desses métodos de otimização e até mesmo na diminuição da quantidade de vezes que o modelo é avaliado. Uma das maneiras de se tratar o problema é a utilização de meta-modelos baseados em métodos de classificação para a avaliação da função de adaptação são capazes de obter boas soluções de maneira mais eficiente. Porém, o que se observa até o momento é que a utilização de meta-modelos raramente é empregado, o que pode ser parcialmente atribuído ao fato de haver uma falta de conhecimento mais amplo da utilização de classificadores lineares ou não-lineares na estimativa de valores da função de ajuste no ambiente de otimização.

Este trabalho tem por objetivo apresentar a utilização de meta-modelos baseados principalmente em classificadores na diminuição do tempo gasto pelos algorítmos genéticos na avaliação da função de ajuste de modelos ambientais. Na seção 2 é discutido o classificador K-NN (*K Nearest Neighbor*, da sigla em inglês) utilizado tanto em algoritmos genéticos mono-objetivos quanto multi-objetivos. Na seção 3, SVM (*Support Vector Machine*, da sigla em inglês) e MSVM (*Multiclass Support Vector Machine*, da sigla em inglês) são abordados e por último, na seção 4, a possibilidade da utilização de alguns outros métodos são apresentadas, árvores de decisão, redes neurais entre outros. Por fim, na seção 5 considerações finais e conclusões são feitas a respeito dos métodos apresentados.

2. Meta-modelos baseados em K-NN

Yang Liu et al (2005) destaca que não importa quão cuidadoso tenha sido o processo de obtenção da função objetivo do problema, esta não é capaz de caracterizar bem todas as peculiaridades dos dados observados, sendo portanto necessário a introdução de algorítmos genéticos multi-objetivo. Neste caso, as avaliações da função de adaptação tornam-se ainda mais complexas. Yang Liu et al (2005) propõe a utilização do método

K-NN (*K Nearest Neighbor*, da sigla em inglês) na obtenção de um meta-modelo que aproxime a função de adaptação sem a necessidade de rodar o modelo numérico para toda iteração do algoritmo genético.

O método K-NN introduzido por E. Fix e J. Hodges (1951) consiste em se definir uma função de distância entre os elementos de um certo conjunto. Usualmente a norma euclidiana é utilizada (1). A partir de então, uma matrix de distâncias é construída considerando todas as possíveis distâncias entre todos os pontos do conjunto. Cada ponto no conjunto é classificado como pertencente à uma classe $C = \{c_1, \dots, c_n\}$. Para se classificar um novo elemento do conjunto considera-se todos os K vizinhos mais próximos dele e contado a quantidade de elementos de cada classe, a classe predominante é então atribuída ao novo elemento.

Considerando K=1, o algoritmo 1-NN parte do princípio de que a partir de uma geração de indivíduos já avaliados segundo a função de adaptação é possível dizer para novos indivíduos qual o valor de sua função de adaptação sem a necessidade de avaliação real da função (rodar o modelo numérico). O classificador pesquisa entre todos os indivíduos qual melhor se assemelha ao indivíduo que se quer classificar (o vizinho mais próximo) e lhe atribui o mesmo valor de função de adaptação. Essa maneira de obter o valor da função de adaptação sem avaliação real da mesma é chamada de avaliação por meta-modelo. O problema passa a ser então decidir quando deve se utilizar a função de adaptação real e quando utilizar o classificador K-NN, para tanto, é necessário a criação de um método de controle de evolução. Dois métodos são propostos por Yang Liu et al (2005).

$$d(X,Y)_{euc} \sqrt{\left(\sum_{i=1}^{n}(X_I - Y_I)^2\right)} \tag{1}$$

(1) **Melhor estratégia:** figura 1. Todos os indivíduos da geração são avaliados segundo o meta-modelo, em seguida são classificados segundo a qualidade de cada um e os *n* melhores indivíduos são avaliados segundo o modelo real.

(2) Alternar entre a avaliação de gerações inteiras hora segundo o meta-modelo, hora segundo o modelo real.

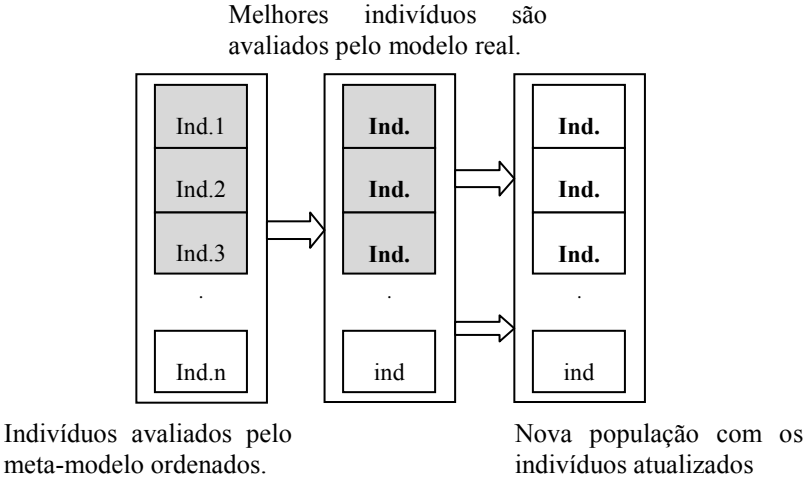

Figura 1. Controle de evolução, melhor estratégia.

Ambas estratégias dependem de vários fatores para se obter resultados satisfatórios, na melhor estratégia, é necessário definir quantos indivíduos serão controlados, já na segunda estratégia, qual a frequência de avaliação de cada modelo, seja ele o real ou o meta-modelo.

Segundo Yang Liu e S. T. Khu (2007) as etapas da construção do algoritmo genetico com o classificador K-NN são as seguinte:

- Inicialização;
- Geração das amostras de treinamento;
- Construção do meta-modelo;
- Avaliar os indivíduos segundo o meta-modelo;
- Ordenar os indivíduos segundo a qualidade;
- Avaliar os melhores indivíduos segundo a função real e atualizar a população;
- Realizar as operações do algoritmo genético;
- Verificar a convergência.

Yang Liu et al (2005) aplica este classificador ao modelo MIKE 11/NAM *rainfall-runoff* H. Madsen (2000) e mostra que foi necessário apneas 38% de avaliação da função de adaptação real, mantendo-se a mesma qualidade de soluções encontradas sem a utilização do classificador.

3. Meta-modelos baseados em SVM e MSVM

Dentre as técnicas de classificação citadas por Yang Liu et al (2005) destaca-se a SVM (*Support Vector Machine*, da sigla em inglês). Trata-se de uma técnica de aprendizado baseada na teoria de aprendizado estatístico Vapnik (1998). Dado um conjunto de dados com *n* amostras (x_i, y_i), onde cada x_i pertencente à $\Re^m$ (conjunto dos números reais de m dimensões) é uma amostra de dado e $y_i \in \{-1, +1\}$, esta técnica procura por um hiperplano ($\mathbf{w} \cdot \mathbf{x} + b = 0$) capaz de separar os dados figura 2. Ou seja, SVM são capazes de fazer a separação dos dados em classes diferentes. Porém, essa abordagem se aplica apenas para casos em que existem 2 classes distintas. Para sua aplicação na classificação de populações nos algoritmos genéticos é necessário que o SVM seja capaz de diferenciar diversas classes, sendo assim, utiliza-se a abordagem por MSVM (*Multiclass Support Vector Machine*, da sigla em inglês). Sendo assim, da mesma maneira como no método K-NN a nova amostra (indivíduo) deve ser classificado quanto à proximidade do indivíduo que apresenta características semelhantes, no MSVM o conjunto de dados da amostra deve ser classificado em classes diferentes e o novo indivíduo será classificado quanto às classes pré-estabelecidas.

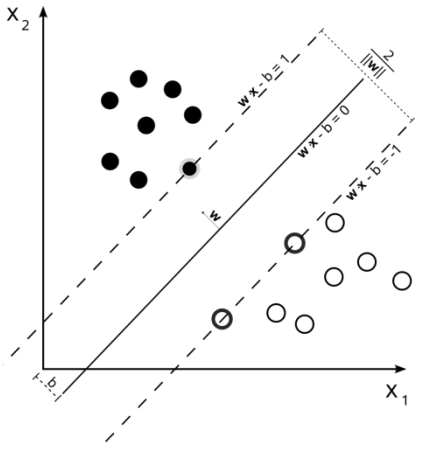

Figura 2. SVM mostrando o hiperplano separador w.x –b=0

4. Outros métodos na construção do meta-modelo

Outros métodos existem como possibilidade na utilização juntamente com algoritmos genéticos na solução do problema, Yang Liu (2005) cita a utilização de redes neurais RBF (*Radial Basis Function*, da sigla em inglês), ou até alguns outros métodos mais populares como aproximação polinomial. J. Bala et al (1995) destaca a utilização de árvores de decisão como mecanismo de avaliação da função de adaptação de algoritmos genéticos no contexto de reconhecimento de padrões. Árvores de decisão é uma estrutura que é capaz de através de aprendizado separar e classificar dados quanto às suas características, sendo assim, funcionaria da mesma maneira que os demais classificadores avaliados, K-NN e MSVM. Além disso, Yaochu Jin (2005) discute várias técnicas de aproximação da função de ajuste em computação evolucionária, método de quadrados mínimos, Modelos de Kriging, *Bagging and boosting* entre outros.

5. Considerações finais e conclusões

A avaliação da qualidade das soluções obtidas na etapa de calibração para modelos ambientais costuma ser um processo muito dispendioso ou até mesmo inviável para alguns modelos. A construção de meta-modelos na avaliação da função de adaptação do algoritmo genético para a calibração de modelos ambientais é proposto neste trabalho como uma maneira de possibilitar que haja uma maior eficiência computacional nesta etapa mantendo-se a mesma qualidade das soluções encontradas caso fosse avaliada a função de adaptação real. Vários métodos são apresentados no qual o meta-modelo pode ser baseado. Segundo Yang Liu e S. T. Khu (2007), um dos principais motivos para a utilização do classificador K-NN é a sua comprovada capacidade como um aproximador de funções, em particular para ser usado na aproximação da função de adaptação do algoritmo genético, além de possuir uma estrutura simples de implementação. Já o classificador MSVM é um método de classificação mais robusto e portanto mais complexo de ser implementado. Apesar de apresentar claramente características que permitam a sua utilização na construção de meta-modelos para tratar o problema e ser citado por Yang liu et al (2005) como uma possibilidade real de aplicação no processo de calibração observa-se que poucos trabalhos Yaochu Jin (2005) existem abordanto esta técnica. Árvores de decisão é mais um exemplo de classificador que pode ser considerado relativamente simples de se implementar e apresenta características para tratar o problema em questão. Outros métodos existem e foram citados por Yaochu Jin

(2005) como por exemplo, aproximação polinomial, rede neural RBF entre outros que são métodos que já são utilizados na matemática como aproximadores de funções e podem portanto serem estudados e a sua viabilidade quanto à aplicação no problema de calibração avaliada. Com o crescimento da complexidade dos modelos cada vez mais se faz necessário a procura por métodos que auxiliem na otimização desses modelos. Além do mais, a maioria dos métodos citados aqui apresentam a possibilidade de paralelismo, o que os coloca em uma condição privilegiada em relação à abordagens por algoritmos sequenciais, já que a computação paralela tem se mostrado bastante promissora com o advento de sistemas multi-processados de baixo custo. Apesar da indiscutível necessidade da utilização de abordagens como as apresentadas por este trabalho, ainda é necessário um grande esforço no intuito de garantir que tais métodos sejam implementados de maneira eficiênciente possibilitando garantir uma maior confiabilidade no processo de modelagem de sistemas ambientais. A aplicação dessas abordagens em problemas práticos de calibração serão investigadas em trabalhos futuros no intuito de permitir uma melhor comparação entre cada metodologia apresentada.

Agradecimentos

Os autores gostariam de agradecer o apoio financeiro providenciado pelo centro nacional de apoio à pesquisas CNPQ.

7. Referências bibliográficas

Yang Liu, C. Zhou, W.J. Ye (2005) "A Fast Optimization Method of Using Nondominated Sorting Genetic Algorithm (NSGA-ll) and 1 -Nearest Neighbor (1 NN) Classifier for Numerical Model Calibration" IEEE.

Yang Liu, S.T. Khu (2007) "Automatic Calibration of Numerical Models Using Fast Optimization by Fitness Approximation" IEEE.

Ana Carolina Lorena, and André C. Ponce de Leon F. de Carvalho (2004) "An Hybrid GA/SVM Approach for Multiclass Classification with Directed Acyclic Graphs – SBIA 2004 pp. 366-375.

J. Bala, J. Huang and H. Vafaie (1995) "Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification" IJCAI 1995.

Yang Liu, Wen-Jing Ye (2005) "Time Consuming Numerical Model Calibration Using Genetic Algorithm (GA), 1-Nearest Neighbor (1NN) Classifier and Principal Component Analysis (PCA)". Proceedings of the 2005 IEEE.

Yaochu Jin (2005) "A Comprehensive Survey of Fitness Approximation in Evolutionary Computation" Soft Computing pp. 3-12.

H. Madsen, (2000) "Automatic Calibration of a Conceptual Rainfall-runoff Model Using Multiple Objectives" Journal of Hydrology pp. 276-288.

E. Fix e J. Hodges (1951) "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties".