ALGORITMOS DE APRENDIZADO DE MÁQUINA APLICADOS À PARÂMETROS MENSURADOS NO RIO ATIBAIA/SP

Maria Rejane Lourençoni Siviero[1] *& Estevam Rafael Hruschka Júnior*[2]

RESUMO - Este artigo trata-se da aplicação de algoritmos de aprendizado máquina: supervisionado e não supervisionado para classificação e agrupamento dos parâmetros mensurados numa seção de medidas no rio Atibaia/SP, visando à previsão da descarga sólida transportada. O banco de dados utilizado compõe-se de quarenta medições, realizadas em uma seção (antiga ponte de trem) no rio Atibaia, contendo os seguintes parâmetros: vazão, declividade da linha d′água, raio hidráulico, largura do espelho d′água, descarga sólida transportada no leito e em suspensão, período 1993 a 1994. Os algoritmos utilizados foram: Árvore de Decisão - C4.5, Naive-Bayes (NB), Regressão Logística (RL) e Expectation Maximization (EM), onde obteve-se os resultados mediante uso do software WEKA (Waikato Environment for Knowledge Analysis) à título de comparação. Os resultados dos algoritmos apresentaram-se com classificação correta e incorreta, respectivamente: C4.5, 40% e 60%; NB, 47,5% e 52,5%; RL, 30% e 70% e EM, 5 agrupamentos: 18%, 18%, 18%, 20% e 28%. Naive-Bayes resultou na melhor classificação, nesse sentido, podendo ser devido ao tamanho do período adotado para as amostras, sugere-se a ampliação do período estudado, bem como nova análise dos resultados para os mesmos algoritmos e aplicação de outros onde a dependência condicional entre os parâmetros seja considerada.

ABSTRACT - This paper presents the applications of machine learning algorithms: supervised and unsupervised of measured parameters in a section Atibaia/SP river to predict load transport. The river data is composed of forty measurements: discharge, water line slope, hydraulic radius, width of the water surface, bed and suspended load transport, period from 1993 to 1994. The algorithms used were: Decision Tree - C4.5, Naive-Bayes (NB), Logistic Regression (LR) and Expectation Maximization (EM). The software WEKA (Waikato Environment for Knowledge Analysis) was using for algorithms and the results to comparison. Thus, the results of the algorithms presented with correct and incorrect classification were respectively: C4.5, 40% and 60% NB, 47.5% and 52.5%, LR, 30% and 70% and EM, 5 groups: 18%, 18%, 18%, 20% and 28%. The of total algorithm used Naive-Bayes was the best result to correct classification in period studied; a suggestion to extend the period studied, as well as new analysis of the results for the same and other algorithms where the conditional dependence between parameters is considered.

Palavras-chave: Aprendizado de máquina na previsão da descarga sólida, aprendizado supervisionado, aprendizado não supervisionado.

1) Pós-Graduação em Ciência da Computação – Universidade Federal de São Carlos, Rod. Washington Luís Km 235, Caixa Postal 676 - 13565905 - São Carlos – SP, e-mail: rsiviero@hotmail.com

2) Professor Associado, Departamento de Computação – Universidade Federal de São Carlos, Rod. Washington Luís Km 235, Caixa Postal 676 - 13565905 - São Carlos – SP, e-mail: estevam@dc.ufscar.br

1 - INTRODUÇÃO

Aprendizado de máquina é um sub-campo da inteligência artificial, dedicado ao desenvolvimento de algoritmos e técnicas que permitam o computador aprender, isto é, que permitam o computador aperfeiçoar seu desempenho em alguma tarefa.

Segundo Russell e Norving (2004) as técnicas de inteligência artificial possuem três características principais: busca (para explorar as distintas possibilidades em problemas onde os passos não são claramente definidos), emprego do conhecimento (permite explorar a estrutura, relações do mundo ou domínio à que pertence o problema e a redução do número de possibilidades a considerar, tal como os humanos fazem) e abstração (proporciona a maneira de generalizar nos passos intrinsecamente similares).

Assim, pode-se utilizar aprendizado de máquina em bancos de dados, reconhecimento de objetos, tais como: face, fala e escrita, diagnósticos médicos, entre outros, porém não se obtém êxito de utilização em sistemas estáticos, como: armazenamento e recuperação de dados.

Embora o aprendizado de máquina seja uma ferramenta poderosa para a aquisição automática de conhecimento, deve ser observado que não existe um único algoritmo que apresente o melhor desempenho para todos os problemas (Monard e Baranauskas, 2005).

Segundo Monard e Baranauskas (2005) é importante compreender o poder e a limitação dos diversos algoritmos de aprendizado de máquina utilizando alguma metodologia que permita avaliar os conceitos induzidos por esses algoritmos em determinados problemas.

De acordo com os conceitos/padrões a serem aprendidos e a disponibilidade de dados para treinamento, pode-se separar em dois tipos de aprendizado, os quais são conhecidos como paradigmas do aprendizado de máquina: aprendizado supervisionado e não supervisionado.

O aprendizado supervisionado possui uma função a ser aprendida, seja de classificação ou de regressão, que está claramente definida, além de o algoritmo conter em sua estrutura uma espécie de instrutor indicando quando a solução está aceitável no exemplo em treinamento e no aprendizado não supervisionado a função a ser aprendida não está explícita e o algoritmo deve aprender os conceitos/padrões, os quais se referem os dados, baseados em agrupamentos e vizinhança.

1.1 - Objetivo

Infelizmente, de forma geral, a ocupação das terras no país se faz de maneira inadequada, pois o uso inadequado do solo vem acelerando o seu depauperamento e, por outro lado, há a contaminação dos corpos d'água pelas partículas de sedimentos, às quais são agentes de absorção, transporte e depósito de pesticidas, compostos orgânicos, bactérias e vírus, entre outros, conduzindo invariavelmente à condições de instabilidade ambiental (Eiger, 2003). Apesar dos esforços que podem ser feitos para mitigar os impactos adversos nos recursos naturais, a carência de dados relacionados aos cursos d'água obtidos por medições e monitoramento ambiental, dificulta sobremaneira esses (Barreto Neto, 2004; Bocarde, 2003).

Dada a importância do rio Atibaia na Bacia do Piracicaba/SP, sendo este o responsável pelo abastecimento de várias comunidades, além de ser o principal receptor das cargas difusas e pontuais da bacia, decidiu-se aplicar algoritmos de aprendizado de máquina em parâmetros mensurados no rio, período de 03/1993 a 12/1994, para classificação e agrupamento dos dados, visando à previsão da descarga sólida transportada.

2 - MATERIAL E MÉTODO

2.1 – Banco de dados

O banco de dados utilizado compõe-se de quarenta medições fluviossedimentométricas realizadas numa seção do rio Atibaia/SP: vazão (Q), declividade da linha d′água (S), raio hidráulico (RH), largura do espelho d′água (B), descarga sólida transportada no leito (Gsb) e em suspensão (Gss), no período de 1993 e 1994 (Siviero, 2003).

2.2 – Software WEKA - Waikato Environment for Knowledge Analysis

O software WEKA é uma coleção de algoritmos de aprendizado de máquina concebido para realizar tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou inseridos a partir do seu próprio código Java. Assim, WEKA possui ferramentas para os dados de pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização (WEKA 3, 2010).

2.3 – Procedimento

O banco de dados original incluia dados de área molhada (A) e perímetro molhado (P), os quais foram retirados, por conter informação redundante, deste modo foi utilizado somente o raio hidráulico (Rh=A/P) e não houve tratamento de valores ausentes, devido os mesmos não existirem.

O banco de dados foi convertido em arquivo.arff, para inserção no banco de dados do WEKA. Foi realizado pré-processamento dos dados para discretização dos dados numéricos utilizando o comando PKI-Discretize. Para evitar overfitting, a validação cruzada (cross-validation) foi escolhida e, após, a seleção dos algoritmos, sendo três de aprendizado supervisionado - C4.5, Naive-Bayes e Regressão Logística e um não supervisionado – EM.

2.3.1 – Árvore de Decisão

É um algoritmo utilizado na tarefa de classificação em aprendizado supervisionado. Trata-se de um algoritmo guloso, o qual faz uma busca ′top-down′ no espaço para todas as possíveis árvores, tendo a entropia (medida da pureza do conjunto de instâncias), utilizada no cálculo da razão de ganho (GR), o qual penaliza os atributos com muitos valores possíveis. Porém, se a amostra for pequena pode ocorrer ′overfitting′ (classificação tendenciosa), uma maneira de evitar o overfitting na árvore de decisão é a poda -′ Occam′s razor′ (Mitechell, 1997).

O aprendizado de árvore de decisão é um dos algoritmos mais utilizados devido a aplicações práticas. É um método de aproximação de funções discretas, robusta a ruídos, capaz de aprender expressões disjuntivas (Mitchell, 1997).

2.3.2 – Naive-Bayes

É um algoritmo utilizado na tarefa de classificação em aprendizado supervisionado, freqüentemente chamado de classificador Bayes (Mitchell, 1997).

Naive Bayes é um modelo probabilístico, sendo um caso particular de rede bayesiana com inferência, onde o número de pais na rede corresponde a um nó; utiliza mecanismo de suavização

de Laplace e a abordagem da probabilidade pode dar-se por máxima verossimilhança ou máxima posteriori.

Segundo Mitchell (2010) é um algoritmo bastante utilizado, seja para variáveis discretas ou contínuas, devido ser de fácil aplicação para treinar com um conjunto de amostras.

Mitchell (1997) afirma: "a performance do naive Bayes em alguns domínios tem mostrado ser comparável com aprendizado de rede neural e a árvore de decisão".

2.3.3 – Regressão Logística

É um algoritmo utilizado na tarefa de classificação em aprendizado supervisionado, geralmente é referido como classificador discriminativo das variáveis.

A regressão logística assume forma paramétrica da distribuição da probabilidade direta, estimando os pesos dos parâmetros dos dados de treinamento; assim, com esse procedimento vai ajustando uma função que define o comportamento desses dados.

Segundo Mitchell (2010) overfitting nos dados de treinamento é um problema que atinge a regressão logística. Assim, uma medida para redução do overfitting é a regularização, cuja função é penalizar grandes valores dos pesos pelo log da máxima verossimilhança.

2.3.4 - Expectation Maximization (EM)

É um algoritmo utilizado nas tarefas de agupamento em aprendizado não supervisionado. Geralmente, ocorre implementação do EM com modelos probabilísticos mais simples, devido às iterações envolvidas no processo até a convergência do agrupamento, dado que agrupamento por definição são processos iterativos, pois não possuem classe definida e necessitam de priori para a inicialização.

3 - RESULTADOS

A Tabela 1, a seguir, contém os resultados dos algoritmos de aprendizado supervisionado e não supervisionado utilizados neste artigo.

Tabela 1 – Resultados obtidos pelos algoritmos

Algoritmo	Classificação	
	Correta	Incorreta
C4.5	16 – 40,0%	24 – 60,0%
Naive-Bayes	19 – 47,5%	21 – 52,5%
Regressão Logística	12 – 30,0%	28 – 70,0%
EM	Agrupamento	
	0 07 (18%)	
	1 07 (18%)	
	2 07 (18%)	
	3 08 (20%)	
	4 11 (28%)	

4 - DISCUSSÃO

A amostra de dados para o treinamento mostrou-se pequena, não refletiu a distribuição dos dados, dado que se traduziu nas classificações obtidas no aprendizado.

O algoritmo de aprendizado de supervisão Naive Bayes foi o que apresentou desempenho melhor, em comparação com Árvore de Decisão C4.5 e Regressão Logística.

O banco de dados foi insuficiente para teste no algoritmo Regressão Logística, pois o mesmo classificou incorretamente 70%, esse algoritmo requer amostras suficientemente grandes.

O algoritmo Expectation Maximization (EM) realizou agrupamento de 5 grupos, nota-se que nos grupos 0, 1 e 2, contém sete elementos em cada com 18%, em um estudo posterior, é preciso verificar quais parâmetros esse algoritmo está agrupando nesses casos.

Supõe-se que, as variáveis do banco de dados, provavelmente, são não lineares e os algoritmos utilizados possuem interação linear, uma possível causa da não obtenção de êxito nas tarefas realizadas pelos algoritmos classificadores.

5 - CONCLUSÃO

Para ter um desempenho melhor nos algoritmos utilizados: Árvore de Decisão - C4.5, Naive-Bayes, Regressão Logística e Expectation Maximization, sugere-se a utilização do banco de dados maior, se disponível, dado que não se conseguiu fazer deduções mais contundentes no período adotado. Por outro, tentar algoritmos com concepções mais elaboradas, no sentido de fazer comparação com outras abordagens, por exemplo, relação de dependência entre os parâmetros.

BIBLIOGRAFIA

BARRETO NETO, A. A. (2004). *Modelagem Dinâmica de Processos Ambientais*. Campinas - SP: Instituto de Geociências - UNICAMP, 123 p. (Tese, doutorado em Geociências – área: Metalogênese).

BOCARDE, F. (2003). *Análise dos Conflitos: Uso e Ocupação da Terra e Fragilidade dos Aquíferos em Paulínia/SP/Brasil*. Campinas - SP: Instituto de Geociências/UNICAMP, 105 p. (Dissertação, mestrado em Geociências – área: Administração e Política de Recursos Minerais).

EIGER, S. (2003). "*Transporte de Poluentes em Meios Aquáticos: Aspectos Conceituai",* in MANCUSO, P. C. S. & SANTOS, H. F. dos (Editores). Reuso de Água. Barueri - SP: Editora Manole Ltda, Cap. 6, p. 175.

MITCHELL, T. (1997). *Machine Learning*. McGraw Hill, New York, 414 p.

MITCHELL, T. (2010). "*Generative and Discriminative classifiers: Naïve Bayes and Logistic Regression"*, in *Machine Learning*. Projeto da 2º ed., Jan. 2010, pp. 1 – 17 (www.cs.cmu.edu/~tom/mlbook.html).

MONARD, M. C.; BARANAUSKAS, J. A. (2005). "*Conceitos sobre Aprendizado de Máquina*", in *Sistemas Inteligêntes: Fundamentos e Aplicações.* Ed. Manole Ltda, Baueri, Cap. 4, pp. 89 – 114.

RUSSELL, S.; NORVING, P. (2004). *Inteligência Artificial*. Editora Campus, 2º ed., Cap. 1, pp. 1 - 31.

SIVIERO, M. R. L. (2003). *Estudo da Ocupação do Solo a Montante de uma Seção do Rio Atibaia Associada à Descarga Sólida Transportada*. Faculdade de Engenharia Civil, Arquitetura e Urbanismo – UNICAMP, 116 p. (Tese, Doutorado em Recursos Hídricos).

WEKA 3 (2010). *Data Mining with Open Source Machine Learning Software in Java*. Universidade de Waikato, Nova Zelândia (http://www.cs.waikato.ac.nz/ml/weka).