# Avaliação de técnicas de mineração de dados para predição de ocorrência do mexilhão dourado em rios da América Latina

**Valter Hugo Guandaline**[1], **Luiz Henrique de Campos Merschmann**[1]

[1]Departamento de Computação - DECOM
Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

vhguandaline@gmail.com, luizhenrique@iceb.ufop.br

***Abstract.*** *The mollusc Limnoperna fortunei (golden mussel) is native to Asia but is spreading around the world mainly transported by transoceanic ships. It is causing several economic and environmental problems due to its greater ability to attach to any solid substratum and its rapid spreading. Therefore, several researchers have published studies related to different aspects of golden mussel. Some of these works used ecological niche models (ENMs) in order to predict the expansion of the invasive golden mussel. However, given the limitations presented by ENMs in previous works, the search for more accurate predictive computational tools is still important. In this work, we evaluated the use of data mining techniques aiming at predicting the occurrence of the golden mussel in Latin America's rivers. The experimental results showed that some classifiers have achieved promising predictive performance.*

***Resumo.*** *O Limnoperna Fortunei (mexilhão dourado) é nativo da Ásia, mas tem se espalhado por todo o mundo principalmente através das águas de lastro dos navios. Devido a sua rápida dispersão e a facilidade de fixação em qualquer superfície, ele tem causado problemas em várias partes do mundo e, por isso, tem chamado atenção de muitos pesquisadores. Vários trabalhos na literatura vêm sendo propostos para investigar diversos aspectos relacionados ao mexilhão dourado. Alguns desses trabalhos abordaram o problema de predição de ocorrência e dispersão do mexilhão utilizando técnicas de modelagem de nicho ecológico. No entanto, dadas as limitações dessas técnicas, a busca por ferramentas preditivas mais precisas torna-se importante. Neste trabalho, avaliamos a utilização de técnicas de mineração de dados para gerar modelos preditivos de ocorrência do mexilhão dourado. Os experimentos realizados mostram que alguns classificadores alcançaram desempenhos preditivos promissores.*

## 1. Introdução

O *Limnoperna Fortunei* [Morton 1975], também conhecido como mexilhão dourado, é um bivalve invasor de origem asiática que tem trazido problemas para o meio ambiente e para a indústria brasileira. Os problemas causados pelo mexilhão dourado são decorrentes da sua facilidade para fixar-se em superfícies de substratos naturais, como rochas e madeiras, e em superfícies de substratos artificiais, tais como cascos de navios, tubulações e outras estruturas construídas por humanos. Esses problemas vão desde a eliminação de espécies nativas até a invasão de estações de tratamento de água e o bloqueio de dutos e turbinas, o que geralmente resulta em grandes impactos econômicos. Para se ter

uma ideia, os problemas ocasionados pelo mexilhão dourado são semelhantes aos gerados pelo mexilhão zebra na América do Norte, onde o custo anual para limpeza das tubulações que foram tomadas pelo mesmo atinge um valor em torno de U$ 5 milhões [Pareschi et al. 2008], [Darrigran et al. 2005].

O primeiro registro do mexilhão dourado na América Latina foi feito em 1991 na costa do Rio de La Plata, província de Buenos Aires, onde foi introduzido acidentalmente através da água de lastro dos navios [Darrigran and Pastorino 1995]. Rapidamente o mexilhão dourado invadiu os principais sistemas aquáticos da Bacia del Plata, expandindo-se para o rio Paraguai atingindo o Pantanal, onde as condições ambientais seriam inóspitas para a maioria dos bivalves [Oliveira et al. 2010]. O mexilhão dourado é encontrado nos rios Paraná e Paraguai e também em rios nos estados de Mato Grosso do Sul, Paraná, São Paulo, Minas Gerais e Rio Grande do Sul [Pestana et al. 2010]. Além disso, registros do mexilhão dourado na costa uruguaia do rio de la Plata e em outros rios do Uruguai, como o rio Santa Lúcia e rio Negro, foram descritos por [Brugnoli et al. 2005].

O mexilhão dourado conseguiu se instalar em diversos rios brasileiros e tem mostrado grande resistência em condições que são consideradas inóspitas para a maioria dos moluscos [Oliveira et al. 2010]. Devido a essa resistência, o mexilhão dourado tem potencial de instalação na maioria dos rios brasileiros e do restante da América Latina. Portanto, a utilização de ferramentas computacionais que auxiliem na predição da invasão e dispersão desse bivalve é fundamental para o planejamento de ações preventivas.

Com base nos conhecimentos sobre os fatores limitantes (depleção de oxigênio diluído, temperatura da água, concentrações de cálcio dissolvido e o carbonato) para o desenvolvimento do mexilhão dourado no Pantanal, em [Oliveira et al. 2010] foram utilizados algoritmos de modelagem de nicho ecológico para predizer o potencial de distribuição do mexilhão dourado nos sistemas fluviais brasileiro e norte-americano. Nesse trabalho, os algoritmos utilizados não conseguiram prever a invasão do mexilhão dourado em vários rios brasileiros onde as concentrações de cálcio e os índices de saturação de calcita são favoráveis ao seu estabelecimento. Segundo os autores, isso indica que os algoritmos utilizados podem ser imprecisos quando os dados utilizados na modelagem não estiverem em faixas de valores similares àquelas das variáveis onde as predições são realizadas.

Dada a importância da predição da invasão e expansão do mexilhão dourado em sistemas fluviais e a limitação dos métodos apresentados na literatura, o objetivo deste trabalho é avaliar a utilização de técnicas de mineração de dados para realizar a predição de ocorrência do mexilhão dourado em rios da América Latina (com foco principal nos rios brasileiros). Mais especificamente, a tarefa de classificação foi empregada para a predição da ocorrência do mexilhão. Para tal, foi construída uma base de dados com características físicas e químicas de diversos pontos onde foram coletadas amostras de água de rios para verificação da presença/ausência do mexilhão dourado.

## 2. Trabalhos Relacionados

Algumas invasões biológicas, além de prejuízos econômicos, podem causar também sérios problemas para o meio ambiente. Devido aos grandes impactos causados pelo mexilhão dourado, vários trabalhos da literatura investigaram diversos aspectos relacionados ao mesmo, tais como características ambientais que propiciam seu estabelecimento, padrão de invasão, velocidade de dispersão e outros. A seguir, são apresentados alguns

trabalhos que abordaram o problema de predição da invasão e dispersão desse bivalve.

Diversas variáveis já foram consideradas em pesquisas cujo foco era estimar o potencial de dispersão de espécies invasoras. Alguns estudos consideraram, por exemplo, os meios de transporte utilizados pela espécie invasora para acessar uma nova localização e as condições que propiciaram a invasão, tais como, os recursos naturais existentes, a presença de inimigos naturais e as características físicas do meio ambiente [Shea and Chesson 2002]. Partindo do princípio de que os meios para introdução de novas espécies estão presentes na maioria dos ambientes, alguns trabalhos realizaram a predição de ocorrência e dispersão de espécies invasoras considerando somente a relação entre características do ambiente receptor da espécie e o nicho ecológico da mesma. Por exemplo, o trabalho [Kluza and McNyset 2005] utilizou modelagem de nicho ecológico para predizer o potencial de dispersão do mexilhão dourado em escala global.

Em [Boltovskoy et al. 2006], os autores realizaram uma análise em escala global sobre os riscos de dispersão do mexilhão dourado em três bacias hidrográficas importantes da América do Sul (do rio Amazonas, do rio Madalena e do rio Orinoco) e, com base em informações relacionadas a existência de caminhos de acesso e mecanismos de transporte para o mexilhão e analisando a adequabilidade das características ambientais para o estabelecimento do mesmo, concluíram que esse bivalve possui grande potencial de dispersão na maioria dos sistemas aquáticos da América do Sul. Eles sugerem também que o mexilhão poderá surgir em rios localizados em áreas que vão desde a América Central até o Canadá. Vale observar que esse trabalho faz apenas uma análise em larga escala sem utilizar qualquer ferramenta computacional para realização das predições.

Visando a predição da ocorrência e dispersão de espécies invasoras numa escala local (em regiões específicas), alguns trabalhos mostraram a importância da utilização de variáveis limnológicas, tais como, concentração de cálcio, pH, temperatura, oxigênio dissolvido, condutividade dentre outras. Nesse sentido, em [Oliveira et al. 2010], a partir do conhecimento sobre os fatores limitantes para a sobrevivência do mexilhão dourado no Pantanal, foram utilizados dois algoritmos de modelagem de nicho ecológico (GARP e MAXENT) para a predição da dispersão do mexilhão dourado em rios do Brasil e da América do Norte. Com relação aos rios brasileiros, os experimentos com os dois algoritmos mostraram que predições corretas foram obtidas em diversos locais e estão de acordo com análises realizadas em [Kluza and McNyset 2005]. No entanto, os autores mencionam que os algorimos falharam na predição da dispersão do mexilhão dourado em rios brasileiros grandes onde a concentração de cálcio e o índice de saturação de calcita são favoráveis ao estabelecimento do mesmo. Eles sugerem que os algoritmos utilizados podem ser imprecisos quando os dados utilizados na modelagem não estiverem em faixas de valores similares àquelas das variáveis onde as predições são realizadas.

## 3. Metodologia

A tarefa de classificação é uma das mais importantes da área de mineração de dados, sendo que seu objetivo é predizer a classe de novas instâncias a partir de uma base de dados contendo instâncias com classes já conhecidas. No caso do problema de predição de ocorrência do mexilhão dourado, as instâncias são compostas por um conjunto de características físicas e químicas de pontos amostrados para verificação da presença ou ausência do mexilhão, cujo resultado corresponde à classe dessas instâncias.

Portanto, para viabilizar a avaliação do uso técnicas de classificação para predição de ocorrência do mexilhão dourado, uma base de dados histórica foi construída a partir de dados coletados na literatura e num repositório disponibilizado pelo CBEIH [1]. Os dados coletados resultaram em uma base de dados com 471 instâncias, sendo 79% relacionadas a registros de presença do mexilhão e 21% de registros de ausência do mesmo. Os dados coletados correspondem ao período de 1991 até 2012.

Dada a diversidade das fontes de dados, o conjunto de características utilizado para descrever cada ponto de coleta de amostra varia de uma fonte de dados para outra. Desse modo, a base de dados construída a partir de diferentes fontes de dados apresenta alto grau de esparsidade. Dentre as características encontradas, nas diferentes fontes, as seguintes foram selecionadas para integrar essa base de dados:

- Características físicas e químicas: temperatura da água, oxigênio dissolvido, condutividade, concentração de cálcio, salinidade, pH, clorofila, turbidez e alcalinidade.
- Características temporais e de localização: latitude, longitude, nome do rio, período de coleta e ano da coleta.
- Registro de presença ou ausência do mexilhão dourado.

A partir da base de dados histórica construída, foram gerados os modelos de classificação para a predição de ocorrência do mexilhão dourado. Vale observar que, apesar de a base de dados possuir informações temporais e de localização geográfica, essas informações não foram utilizadas na construção dos classificadores. Sendo assim, apenas as características físicas e químicas (além do atributo classe) foram utilizadas no processo de classificação.

Em mineração de dados, o processo de classificação é dividido em duas etapas: a etapa de treinamento do modelo de classificação e, depois, a etapa de teste (avaliação) do mesmo. O objetivo da etapa de treinamento é construir um modelo de classificação a partir de um subconjunto de instâncias contidas na base de dados histórica. Cada instância dessa base de dados é caracterizada por um conjunto de atributos preditores e pertence a uma determinada classe. As instâncias utilizadas para construção do modelo de classificação formam a base de dados de treinamento. Na etapa de teste utilizam-se as instâncias da base de dados histórica que não foram consideradas na fase de treinamento. Esse conjunto de instâncias compõe a base de dados de teste. Nessa etapa, o modelo gerado na fase de treinamento é avaliado utilizando-se somente as instâncias contidas na base de teste.

Assim como já mencionado, a base de dados histórica construída para treinamento e teste dos modelos de classificação é esparsa, ou seja, 86,5% das instâncias possuem pelo menos 44% dos atributos (características físicas e químicas) com valores desconhecidos. Dado que valores ausentes de atributos podem prejudicar o desempenho preditivo dos classificadores, um pré-processamento dessa base de dados foi realizado com o objetivo de preencher esses valores ausentes. Neste trabalho, foram avaliadas três diferentes estratégias para tratar o problema de ausência de valores de atributos.

## 4. Experimentos

Dentre os vários métodos de classificação propostos na literatura, neste trabalho foram utilizados o Naive Bayes [Duda and Hart 1973], $k$-NN [Aha and Kibler 1991] e o Ran-

[1] Centro de Bioengenharia de Espécies Invasoras de Hidrelétricas. Disponível em http://www.cbeih.org/.

dom Forest [Breiman 2001]. Esses classificadores estão implementados na ferramenta Weka[2] (v.3.7) e trabalham com bases de dados contendo valores ausentes de atributos.

Diferentes valores de parâmetros foram avaliados tanto para o $k$-NN como para o Random Forest. Para o $k$-NN foram avaliados $k$ = 1, 3, 5 e 10. O melhor resultado foi obtido com $k$ = 5 e, portanto, esse foi o valor adotado nos experimentos aqui relatados. No caso do Random Forest, os seguintes valores foram avaliados para o parâmetro referente ao número de árvores ($numTrees$): 1, 5, 10, 15 e 20. Nesse caso, $numTrees = 10$ alcançou o melhor desempenho preditivo e, por isso, foi adotado para realização dos experimentos. Já o parâmetro relacionado com a seleção aleatória de atributos ($numFeatures$) é calculado a partir da equação $log_2(n) + 1$, onde $n$ corresponde ao número de atributos preditores da base de dados. No caso da base de dados utilizada neste trabalho, $n = 9$.

Na implementação do Weka os classificadores Naive Bayes e Random Forest tratam os valores ausentes da mesma maneira, isto é, tanto na etapa de treinamento quanto na etapa teste, esses classificadores ignoram os atributos com valores ausentes. Já o $k$-NN, ao realizar o cálculo da distância entre as instâncias, se o valor de um atributo de pelo menos uma das instâncias estiver faltando, ele considera a distância desse atributo entre essas instâncias como máxima.

Para avaliação dos classificadores foi utilizado o método 10-validação cruzada. Além disso, duas categorias de base de dados de treinamento foram avaliadas. A primeira, denominada base completa, contém todas as instâncias de treinamento da base original. Já a segunda, denominada base resumida, contém apenas as instâncias de treinamento da base de dados original que contém pelo menos 50% dos valores dos atributos presentes. Nas duas categorias existe um desbalanceamento entre a quantidade de instâncias pertencentes às classes "sim" (presença do mexilhão) e "não" (ausência do mexilhão). Enquanto na base completa 79% das instâncias são da classe "sim" e 21% são da classe "não", na base resumida esses percentuais mudam para 42% e 58%, respectivamente. Já a base de dados de teste é sempre a mesma para propiciar uma comparação de desempenho justa quando se utiliza cada uma das bases de treinamento (completa e resumida).

Para as bases de treinamento completa e resumida foi realizado um pré-processamento para substituir os valores ausentes de atributos. Desse modo, novas bases de treinamento foram obtidas a partir da utilização das seguintes técnicas: substituição de valores ausentes pelo valor médio do atributo (MG e MG2), substituição de valores ausentes pelo valor médio do atributo em cada classe (MPC) e substituição de valores ausentes utilizando-se a técnica denominada Expectation Maximization (EM)[Schafer 1997].

A diferença entre as estratégias aqui denominadas MG e MG2 consiste no fato de que enquanto na estratégia MG2 são substituídos os valores ausentes apenas das instâncias de treinamento, na estratégia MG, a substituição de valores ausentes também acontece nas instâncias de teste (onde o valor substituído também corresponde ao valor médio do atributo calculado a partir da base de treino).

Na estratégia de substituição pela média de cada classe (MPC) são calculadas duas médias para cada atributo, uma considerando os valores do atributo das instâncias pertencentes à classe "sim" e outra a partir dos valores do atributo das instâncias da classe

[2]Disponivel em http://www.cs.waikato.ac.nz/ml/weka/

"não". Desse modo, cada instância de treinamento contendo valores ausentes de atributos recebe os valores médios calculados para sua respectiva classe. Por exemplo, se a média do atributo "Temperatura" calculado para a classe "sim" for de 15 °C, todas as instâncias que não possuem valor para o atributo "Temperatura" e pertencem à classe "sim" passarão a ter o valor 15 °C para esse atributo.

A técnica EM corresponde a um procedimento iterativo que realiza uma estimativa dos parâmetros da função de distribuição estatística de uma amostra (atributo) e, a partir dessa função, realiza a predição dos valores ausentes do atributo [Schafer 1997].

Os principais objetivos desses experimentos são: $a$) Verificar se a remoção de instâncias de treinamento com mais de 50% de valores ausentes de atributos (base resumida) melhora o desempenho preditivo dos classificadores; $b$) Avaliar a influência das abordagens de substituição de valores ausentes de atributos no desempenho dos classificadores; $c$) Comparar o desempenho preditivo dos classificadores avaliados neste trabalho.

## 5. Resultados

A medida utilizada para avaliar o desempenho dos classificadores para cada base de dados (completa e resumida) foi a F-Measure. Os gráficos a seguir apresentam o desempenho dos classificadores para as bases completa e resumida considerando-se as diferentes abordagens de substituição de valores ausentes de atributos (representadas no eixo X), onde "SP" (sem pré-processamento) significa que a base de dados não sofreu qualquer tratamento de substituição de valores ausentes de atributos. Nos demais casos, as siglas representam as abordagens de substituição apresentadas na Seção 4.

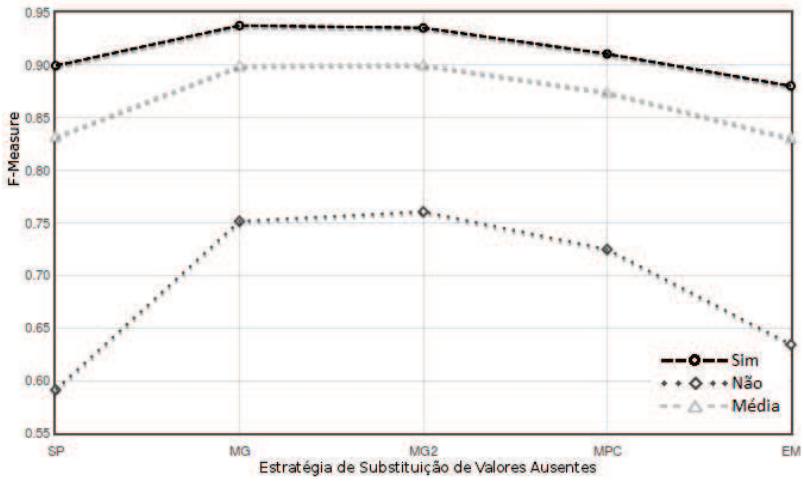


**Figura 1. F-Measure do Naive Bayes para base completa**

A Figura 1 apresenta o gráfico de desempenho do classificador Naive Bayes para a base completa com as diferentes abordagens de substituição de valores ausentes. Nessa base, o Naive Bayes apresentou um bom desempenho ao classificar as instâncias da classe "sim", com F-measure variando entre 88% (EM) e 93,8% (MG). Porém, seu desempenho com as instâncias da classe "não" não seguiu o mesmo padrão, obtendo valores de F-measure entre 59,1%(SP) e 76,1%(MG2). Além disso, pode-se observar também que, exceto para a técnica EM, pré-processar a base de dados substituindo os valores ausentes de atributos ajudou a melhorar o desempenho desse classificador. Vale observar que a

linha que está entre aquelas que representam as classes “sim” e “não” corresponde à média ponderada do F-measure entre essas duas classes.

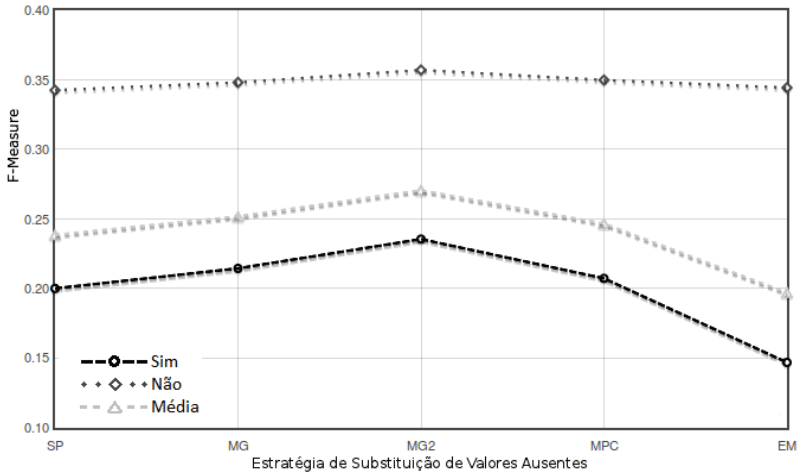


**Figura 2. F-Measure do Naive Bayes para base resumida**

A Figura 2 mostra o desempenho do Naive Bayes para a base de dados resumida. Dada a alteração da distribuição de classes quando comparada com a base completa, nessa base, independentemente da estratégia de substituição de valores ausentes, o classificador alcançou melhor desempenho para as instâncias da classe “não”.

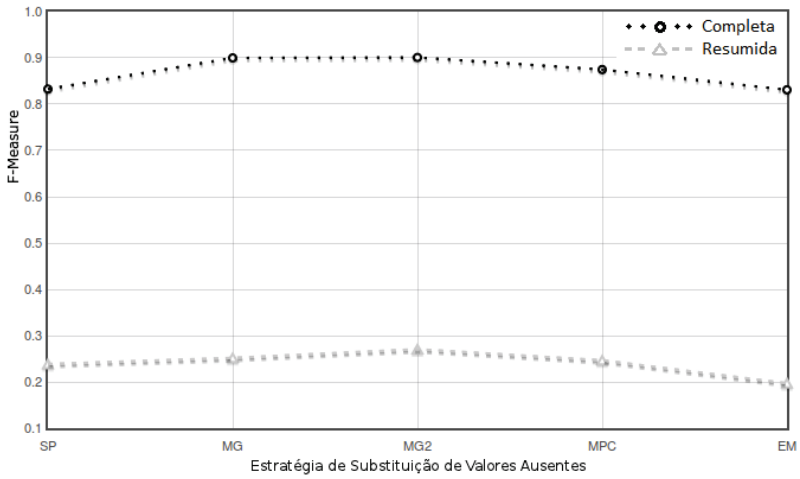


**Figura 3. F-Measure do Naive Bayes para as bases completa e resumida**

A Figura 3 mostra o F-measure médio obtido pelo Naive Bayes para as bases completa e resumida. Pode-se notar que o desempenho do classificador para a base resumida foi muito pior do que para a base completa, indicando que remover instâncias com muitos valores ausentes de atributos não foi uma boa estratégia para esse classificador.

O gráfico da Figura 4 mostra o desempenho do algoritmo $k$-NN para base completa. Sem aplicar um pré-processamento (SP) o $K$-NN teve um baixo desempenho ao classificar instâncias de ambas as classes (11,3% para a classe “sim” e 37,4% para a classe “não”). O fato de o F-measure ter sido maior para as instâncias da classe “não” pode estar

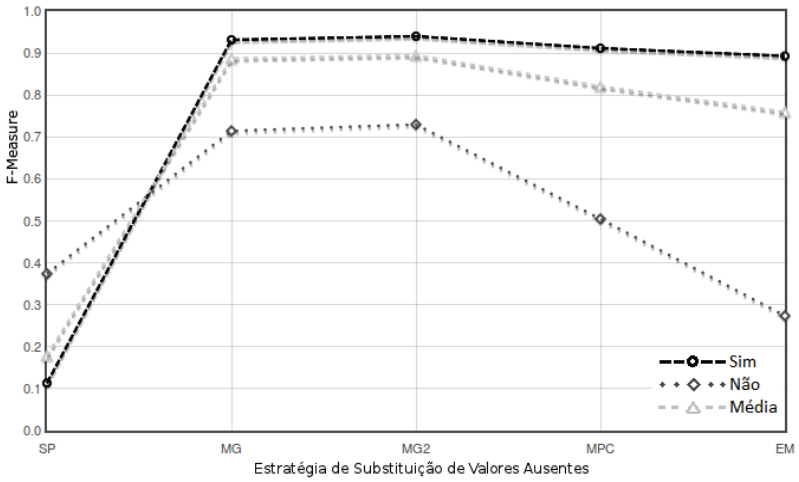


**Figura 4. F-Measure do Knn para a base completa**

relacionado com o fato de que, apesar de a base completa ter menos instâncias da classe "não" do que da classe "sim", as instâncias da classe "não" estão mais completas do que as da classe "sim". Por outro lado, para todas as abordagens de substituição de valores ausentes avaliadas, o $k$-NN passou a ter desempenho melhor para as instâncias da classe "sim". Além disso, verifica-se que o desempenho médio desse classificador para todas as abordagens de substituição de valores ausentes é melhor do que aquele obtido a partir da base sem esse pré-processamento.

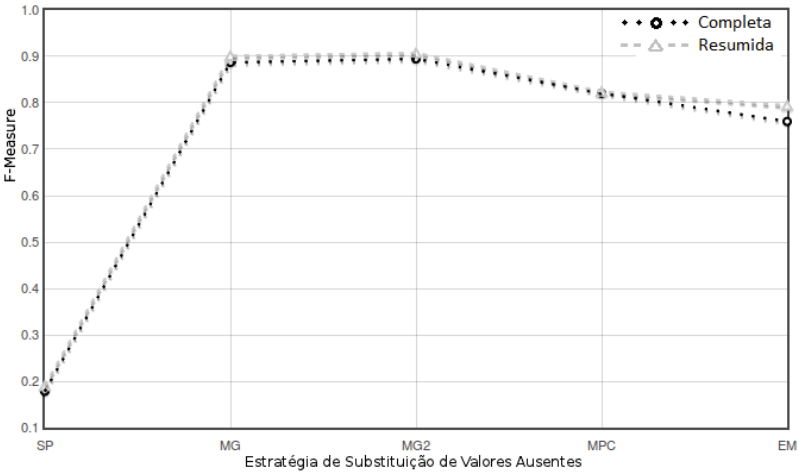


**Figura 5. F-Measure do Knn para as bases completa e resumida**

A Figura 5 apresenta o F-measure médio alcançado pelo $k$-NN para as bases completa e resumida. Para a base resumida, o desempenho do $k$-NN foi muito semelhante àquele alcançado com a base completa, o que indica que a redução do número de instâncias da base não afetou o seu poder preditivo.

O gráfico da Figura 6 mostra o desempenho do Random Forest para cada uma das classes da base de dados completa. Assim como na maioria dos casos apresentados até aqui, pode-se observar que o F-Measure é maior para as instâncias pertencentes à classe "sim". Um dos motivos da superioridade de desempenho para instâncias da classe

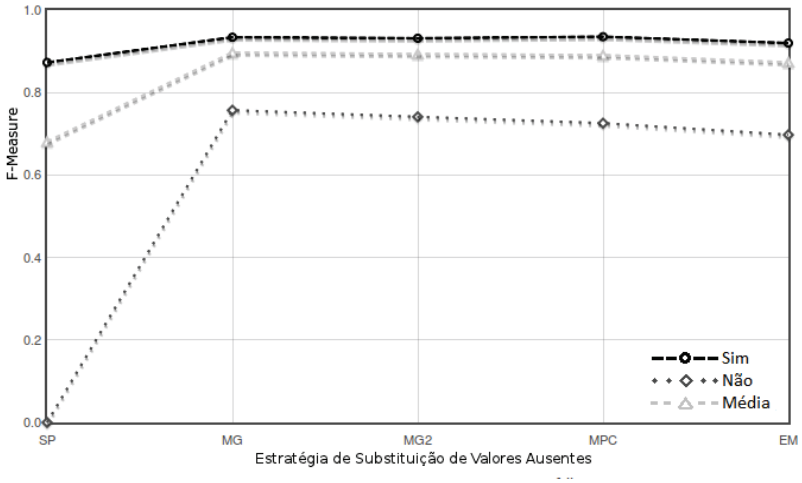


**Figura 6. F-Measure do Random Forest para a base completa**

"sim" está relacionado com o grande desbalanceamento das classes na base de dados completa. Além disso, verifica-se que realizar um pré-processamento para substituir os valores ausentes de atributos, independentemente da técnica adotada, é sempre melhor do que trabalhar com a base sem pré-processamento (SP). Aliás, no caso desse classificador, vale observar que para a base de dados onde não foi feita a substituição de valores ausentes (SP), ele classificou erroneamente todas as instância da classe "não" (F-measure = 0%).

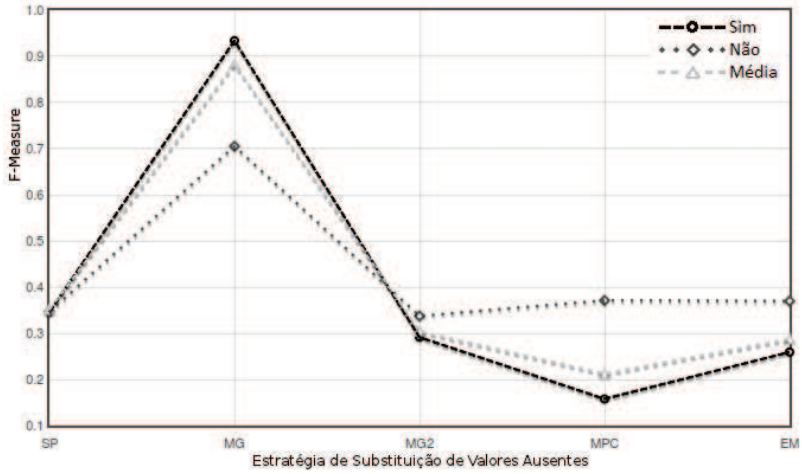


**Figura 7. F-Measure do Random Forest para a base resumida**

O gráfico da Figura 7 apresenta o desempenho do Random Forest para a base de dados resumida. Pode-se observar que o Random Forest apresentou superioridade na predição das instâncias pertencentes à classe "sim" somente para a estratégia de substituição MG, passando a ter desempenho superior para as instâncias da classe "não" com as técnicas MG2, MPC e EM. O bom desempenho do Random Forest com a base resumida para a estratégia MG (F-measure = 88,4%) deve estar relacionado com o fato de essa estratégia também preencher os valores ausentes das instâncias de teste.

Ao se comparar o desempenho do Random Forest nas bases completa e resumida, podemos concluir a partir do gráfico da Figura 8 que, exceto para a estratégia MG (onde

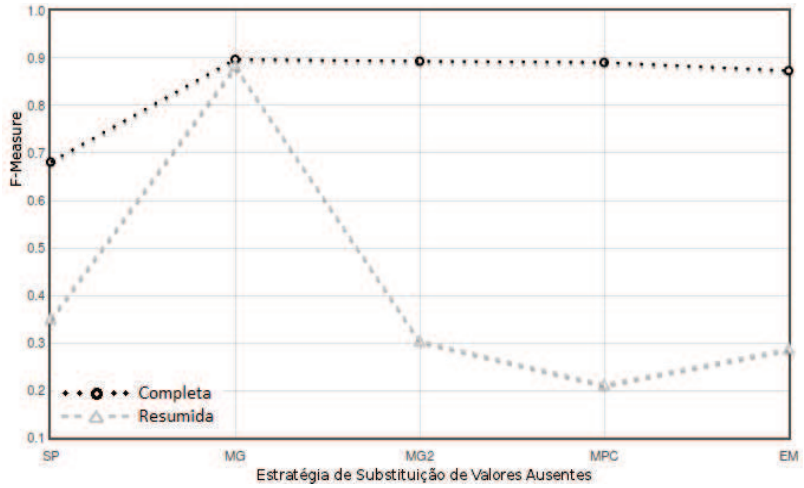


**Figura 8. F-Measure do Random Forest para as bases completa e resumida**

os desempenhos foram equivalentes), a redução da base de dados trouxe degradação de desempenho para esse classificador.

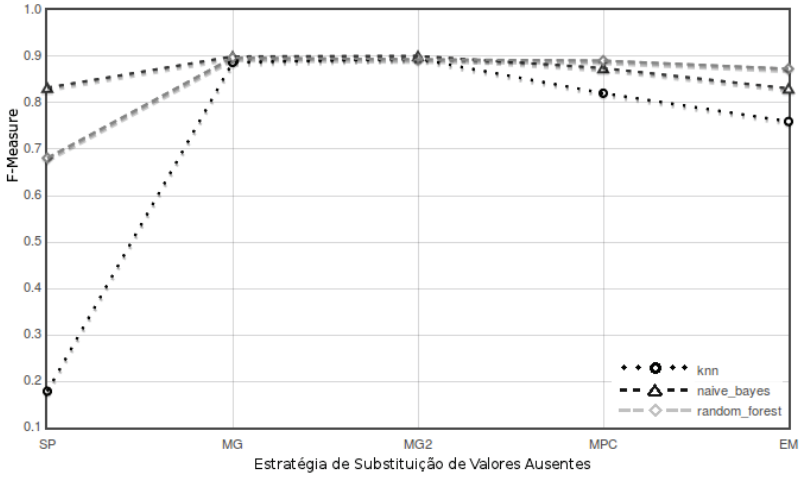


**Figura 9. F-Measure dos classificadores para a base completa**

Para finalizar as análises comparativas, a Figura 9 mostra o gráfico de desempenho médio de todos os classificadores para a base completa, com a qual sempre obtivemos superioridade de desempenho. Nesse gráfico é possível notar que todos os classificadores apresentaram melhoria de desempenho quando houve algum tipo de tratamento para o problema dos valores ausentes de atributos. Além disso, o Naive Bayes e o Random Forest foram os classificadores que apresentaram os melhores desempenhos, os quais estão relacionados com as estratégias de substituição MG e MG2.

## 6. Conclusão

Neste trabalho avaliamos o uso de técnicas de mineração de dados, mais especificamente técnicas de classificação, para gerar modelos capazes de predizer a ocorrência do mexilhão dourado em diversos pontos de rios da América Latina. Para isso, construímos uma base de dados histórica com registros de presença e ausência do mexilhão dourado.

Dada a diversidade das fontes de dados utilizadas para coleta dos registros, nossa base de dados histórica apresentou alto grau de esparsidade, ou seja, 86,5% das instâncias da base possuem pelo menos 44% dos atributos cujos valores são desconhecidos. Dado que o desempenho dos classificadores pode ser negativamente afetado por essa ausência de valores de atributos, diferentes estratégias de substituição de valores ausentes foram avaliadas na etapa de pré-processamento da base. Os classificadores avaliados neste trabalho foram o Naive Bayes, o Random Forest e o $k$-NN.

Experimentos também foram conduzidos com uma base de dados (resumida) que foi construída removendo-se da base original todas as instâncias que apresentavam mais de 50% de seus atributos com valores ausentes. O objetivo desses experimentos foi verificar o impacto da eliminação das instâncias mais incompletas (com mais valores ausentes de atributos) no desempenho preditivo dos classificadores. Os resultados apresentados na Seção 5 mostraram que, apesar de a eliminação dessas instâncias tornar a base de dados menos esparsa, independentemente da realização ou não de pré-processamento para substituição de valores ausentes, para nenhum classificador essa eliminação de instâncias resultou em um desempenho melhor do que aquele obtido com a base completa.

Para a base completa, os resultados experimentais mostraram que os classificadores Random Forest e Naive Bayes foram os que apresentaram os melhores desempenhos (em termos de F-measure médio). Além disso, esses melhores desempenhos foram obtidos para as estratégias de substituição de valores ausentes denominadas MG (90% com Naive Bayes e 89,7% com Random Forest) e MG2 (90,1% com Naive Bayes e 89,4% com Random Forest).

Apesar de outros trabalhos na literatura já terem abordado o problema da predição de ocorrência do mexilhão dourado, os resultados apresentados por esses trabalhos não foram comparados com aqueles aqui obtidos pelos motivos descritos a seguir. Em primeiro lugar, por não termos acesso aos dados utilizados nesses trabalhos anteriores, construímos e utilizamos no treinamento dos modelos de classificação uma base de dados própria, o que nos permite afirmar que não há garantia de que os dados aqui utilizados sejam os mesmos dos trabalhos apresentados na literatura. Além disso, não há um consenso entre as variáveis limnológicas consideradas nos trabalhos anteriores e as utilizadas na indução dos modelos preditivos deste trabalho, onde tentamos adotar o maior número possível de variáveis que, segundo a literatura, estão relacionadas com a ocorrência e dispersão do mexilhão dourado. No entanto, podemos considerar que os resultados obtidos neste trabalho são satisfatórios (com desempenho em termos de F-measure médio em torno de 90%), o que indica a viabilidade da aplicação de técnicas de mineração de dados para solucionar o problema em questão.

Vale ressaltar que, apesar de os classificadores Random Forest e Naive Bayes terem alcançado F-measure médio de torno de 90%, devido ao desbalanceamento de classes da base completa, os desempenhos preditivos para as instâncias pertencentes à classe "sim" foram sempre superiores àqueles obtidos para as instâncias da classe "não". Desse modo, um trabalho futuro será avaliar técnicas de balanceamento de classes na tentativa de melhorar o desempenho preditivo dos classificadores para as instâncias da classe "não". Além disso, estratégias de substituição de valores ausentes de atributos baseadas em informações de localização geográfica e temporais serão propostas e comparadas com aquelas que apresentaram os melhores resultados nos experimentos aqui realizados.

## Agradecimentos

Este trabalho é resultado do projeto Rede REAlf – Rede de Estudos Avançados em Limnoperna fortunei, financiado pela FAPEMIG e pelo Instituto Tecnológico VALE.

## Referências

Aha, D. and Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Boltovskoy, D., Correa, N., Cataldo, D., and Sylvester, F. (2006). Dispersion and ecological impact of the invasive freshwater bivalve limnoperna fortunei in the río de la plata watershed and beyond. *Biological Invasions*, 8(4):947–963.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brugnoli, E., Clemente, J., Boccardi, L., Borthagaray, A., and Scarabino, F. (2005). Golden mussel limnoperna fortunei (bivalvia: Mytilidae) distribution in the main hydrographical basins of uruguay: update and predictions. *Anais da Academia Brasileira de Ciências*, 77(2):235–244.

Darrigran, G., Damborenea, M., and Penchaszadeh, P. (2005). El mejillón dorado limnoperna fortunei (dunker, 1857) en la cuenca del plata. *Invasores: Invertebrados exóticos en el Río de la Plata y región marina aledaña. Buenos Aires: Eudeba*, pages 39–102.

Darrigran, G. and Pastorino, G. (1995). The recent introduction of a freshwater asiatic bivalve, limnoperna fortunei (mytilidae) into south america. *Veliger*, 38(2):171–175.

Duda, R. and Hart, P. (1973). Pattern classification and scene analysis.

Kluza, D. and McNyset, K. (2005). Ecological niche modeling of aquatic invasive species. *Aquatic Invaders*, 16:1–7.

Morton, B. (1975). The colonization of hong kong's raw water supply system by limnoperna fortunei (dunker 1857)(bivalvia: Mytilacea) from china. *Malacol. Rev.*, 8:91–105.

Oliveira, M. D., Hamilton, S. K., and Jacobi, C. M. (2010). Forecasting the expansion of the invasive golden mussel limnopernafortunei in brazilian and north american rivers based on its occurrence in the paraguay river and pantanal wetland of brazil. *Aquatic Invasions*, 5(1):59–73.

Pareschi, D., Matsumura-Tundisi, T., Medeiros, G., Luzia, A., and Tundisi, J. (2008). First occurrence of limnoperna fortunei (dunker, 1857) in the rio tietê watershed (são paulo state, brazil). *Brazilian Journal of Biology*, 68(4):1107–1114.

Pestana, D., Ostrensky, A., Tschá, M. K., and Boeger, W. A. (2010). Prospecção do molusco invasor limnoperna fortunei (dunker, 1857) nos principais corpos hídricos do estado do paraná, brasil. *Papéis Avulsos de Zoologia (São Paulo)*, 50(34):553–559.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*.

Shea, K. and Chesson, P. (2002). Community ecology theory as a framework for biological invasions. *Trends in Ecology & Evolution*, 17(4):170 – 176.