SCIENTIA PLENA
www.scientiaplena.org.br
VOL. 11, NUM. 08 2015
doi: 10.14808/sci.plena.2015.081326

Extração de conhecimento a partir de bancos de dados oceanográficos mistos

Knowledge discovery using mixed oceanographic data

M. M. Barbat[1*]; J. L. Coletto[2]; M. G. Goulart[1]; E. P. Teixeira[1]; K. Machado[1]; E. N. Borges[1]

[1]*Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande-RS, Brasil*

2 Instituto Oceanográfico – Universidade Federal do Rio Grande, Rio Grande-RS, Brasil

**maurobarbat@gmail.com*

(Recebido em 22 de setembro de 2014; aceito em 29 de dezembro de 2014)

Em estudos oceanográficos é comum o uso de diferentes bases de dados visando correlacionar estruturas comuns que evidenciem ou comprovem determinado aspecto. Tal como dados oceanográficos e meteorológicos provenientes de satélites, boias, entre outros, disponibilizados publicamente por órgãos internacionais de pesquisa. O objetivo deste trabalho consiste em aplicar um processo de extração de conhecimento de forma a correlacionar dados de sensoriamento remoto como temperatura do oceano e níveis de clorofila-a, com dados pontuais de captura provenientes de empresas de pesca. Desta forma, a abundância do recurso vivo Bonito-Listrado (Katsuwonus Pelamis) foi correlacionada com parâmetros ambientais provenientes de sensores satelitais.

Palavras-chave: descoberta de conhecimento; mineração de dados; oceanografia.

In oceanographic studies is common the use of different data bases in order to correlate common structures that show evidence of a certain aspect. Like oceanographic and meteorological data obtained through satellites probes and another made available by international research corporations. This paper aims at applying a knowledge extraction process in order to correlate data from remote sensing like ocean temperature and chlorophyll-a levels with fish catch data from fishing companies. More specifically, to correlate the abundance of live resource skipjack tuna (Katsuwonus Pelamis) with these environmental parameters from satellite sensors.

Keywords: Knowledge discovery; data mining; oceanography.

1. INTRODUÇÃO

O uso de técnicas e ferramentas que auxiliem o processo de obtenção de conhecimento, através da exploração de bases de dados, é essencial. Bancos de dados complexos podem tornar a análise de forma manual impraticável ou sujeita a falhas.

De forma geral, o processo de extração do conhecimento proposto por Fayyad [2] pode ser descrito como mostrado na Figura 1, composto por cinco etapas principais: seleção, pré-processamento, formatação, mineração de dados e avaliação.

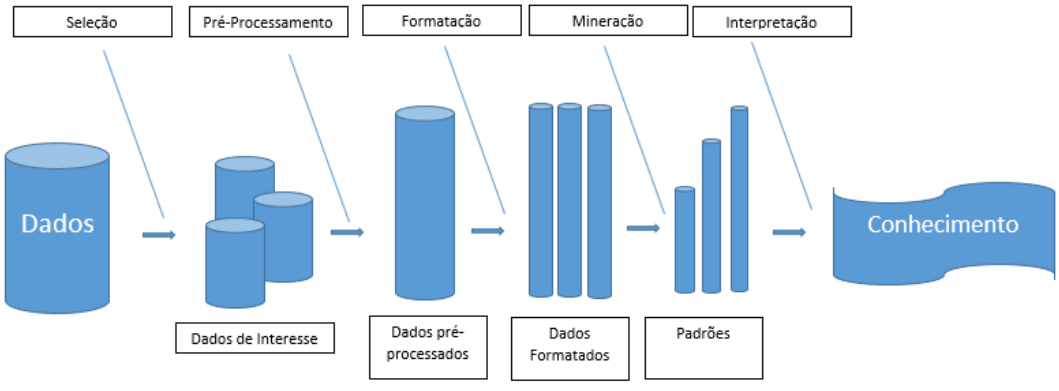

Figura 1: Passos da descoberta de conhecimento proposto por Fayyad [2]

O objetivo deste trabalho consiste em aplicar um processo de descoberta de conhecimento (do inglês Knowledge Discovery in Databases - KDD) para relacionar dados ambientais de sensoriamento remoto, relativos à temperatura da superfície do mar e à clorofila a, com dados de captura do atum Bonito Listrado (Katsuwonus pelamis). Este processo é realizado na tentativa de identificar padrões ambientais que possam estar associados aos processos oceanográficos que influenciam as ocorrências desta espécie em determinados setores e com efeito nas capturas nas áreas de pesca.

O artigo organiza-se da seguinte forma: a Seção 2 descreve os materiais e métodos utilizados neste trabalho como o processo de KDD e os dados utilizados. A Seção 3 apresenta os resultados obtidos incluindo os modelos de árvores de decisão e a avaliação dos mesmos. Por fim, a Seção 4 descreve as conclusões do trabalho.

2. MATERIAL E MÉTODOS

2.1 Dados utilizados e pré-processamento.

A primeira etapa da descoberta de conhecimento consiste no pré-processamento dos dados os quais neste trabalho proposto são constituídos essencialmente de séries temporais, obtidos de diferentes bases:

- Dados de Temperatura da Superfície do Mar (TSM) e Clorofila-a (CHLA), fornecidos pela NOAA NASA[1];
- Dados de taxas de captura, fornecido por empresa de processamento de pescado;
- Dados de posição global de captura;
- Dados de profundidade do local de captura.

Todos os dados obtidos são georreferenciados uma vez que o estudo é realizado na área de captura do atum, de forma que a distribuição espacial e temporal dos dados é importante.

Os dados brutos de sensoriamento remoto são disponibilizados na forma de imagens globais (Figura 2) com resolução espacial de 4 km e sujeitas a falhas ocasionadas por presença de nuvens que geram ausência de informação em determinados pontos que podem ocorrer na área de interesse. No total foram obtidas 256 posições de captura de atum e 198 imagens de sensoriamento remoto, totalizando um montante de 16.8 GB de dados a serem pré-processados.

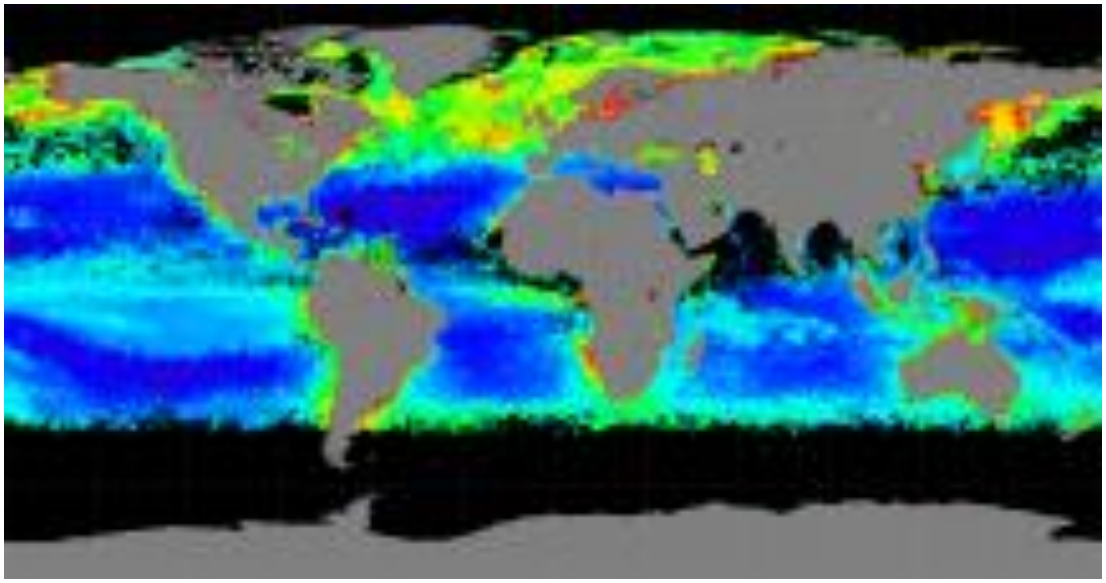

Figura 2: Dados de exemplo – Concentração de Clorofila A (mg/m³)

[1] www.noaa.gov

A etapa subsequente foi o processamento e extração dos valores numéricos das variáveis ambientais quando se executou a extração dos valores médios de TSM e CHLA, nas áreas de captura. Este trabalho foi realizado através de analise estatística em nível de pixel, em áreas de 1 mn^2 (milha náutica quadrada aproximadamente 3,43 km^2) com centro nos locais de captura do pescado, conforme Figuras 3 e 4.

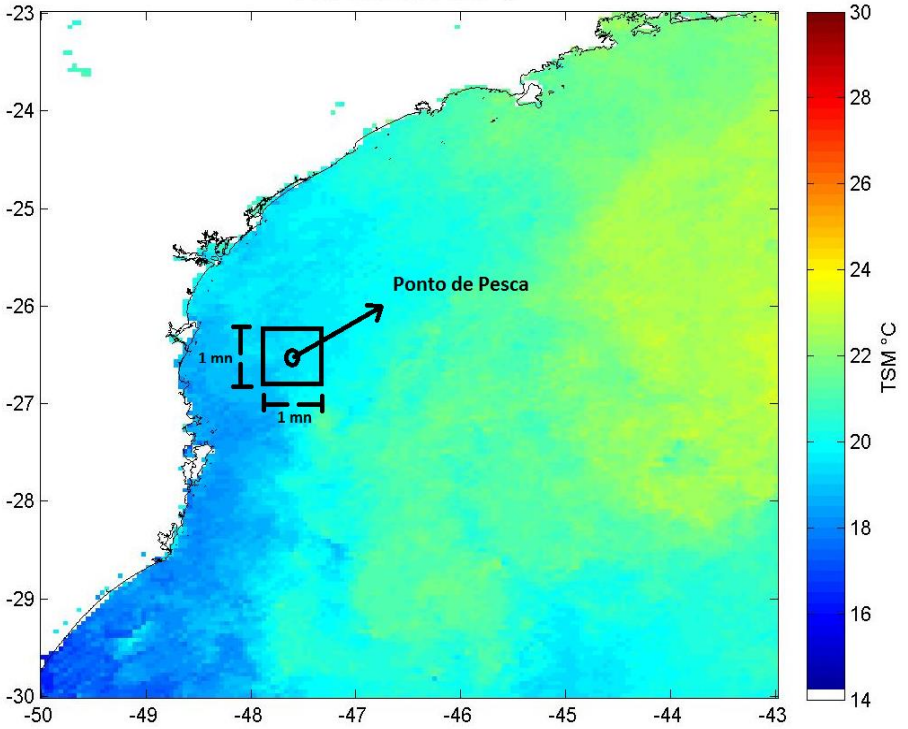

Figura 3: Exemplo da extração dos dados estatísticos de uma área

Na sequência foram extraídas a média e variância da temperatura e clorofila referentes às áreas que contém as capturas. Estas características ambientais podem influenciar na presença do bonito, bem como nos seus processos de alimentação e migração, como citado por [1].

.

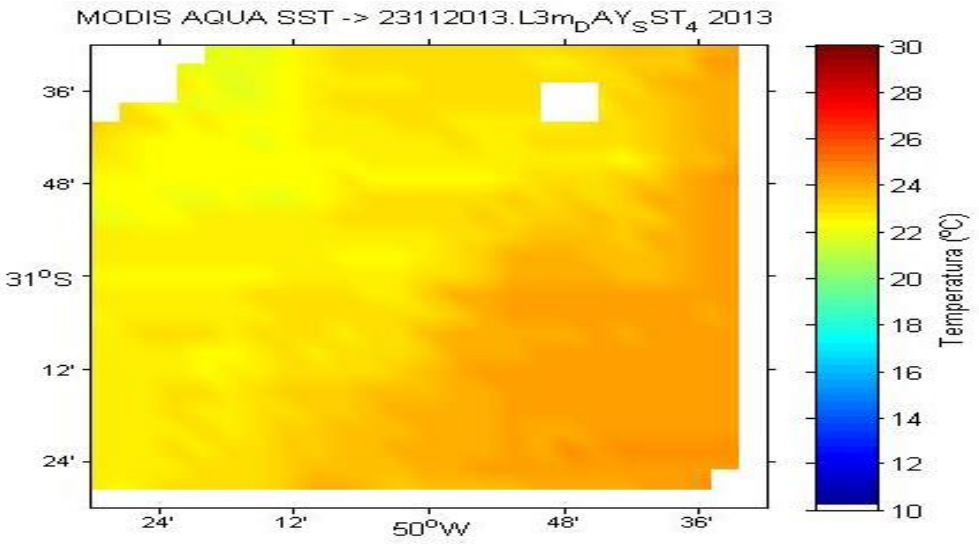

Figura 4: Área de captura

Após as etapas de pré-processamento foi obtido uma base de dados formatada, contendo as informações pertinentes às etapas seguintes, visando correlacionar características ambientais à ocorrência do pescado.

A base de dados resultante é composta por 8 atributos e 256 instâncias, conforme apresentado na amostra de duas instancias na Tabela 1. Os atributos resultantes são referentes a posição global do ponto de pesca (Latitude/Longitude), a temperatura média da superfície do mar (MSST), variância média da temperatura da superfície do mar (VSST), média da concentração de clorofila-a (MCHLA), variância da concentração média de clorofila-a (VCHLA), profundidade referente ao ponto de captura do pescado e o massa total de captura, respectivamente.

Tabela 1. Exemplo da Base de dados formatada para a aplicação da mineração de dados.

Latitude	Longitude	MSST	VSST	MCHLA	VCHLA	Prof	kg
32,92	50,49	23,31	0,09	0,6	0,025	125	1537,5
31,31	49,6	23,2	0,47	0,425	0,017	139	2511,67

2.2 Mineração de dados

Como o objetivo do trabalho é relacionar as características ambientais com a quantidade de captura do Atum Bonito Listrado, o atributo alvo inicialmente era de natureza contínua. Sendo assim, entre as diferentes tarefas de mineração de dados decidiu-se utilizar a tarefa de regressão [3], sendo esta uma técnica de mineração de dados (aprendizagem de máquina) usada para ajustar uma equação para um conjunto de dados

Devido à natureza contínua do atributo alvo (pesca em Kilogramas), como primeiro algoritmo de mineração de dados foi aplicado árvores de regressão (ou árvore modelo) (Figura 5) implementadas através do algoritmo M5 [4]. Este algoritmo gera uma árvore de modelos cujos nós folhas são funções lineares, capazes de predizer o valor do atributo-meta, neste caso, a captura total. O algoritmo utilizado está implementado na ferramenta Weka [5].

Quando se utiliza uma árvore de modelos para predizer o valor de uma determinada instância percorre-se a árvore da raiz às folhas, utilizando os valores dos atributos das instâncias para a tomada de decisão.

Os dados foram analisados segundo duas metodologias, buscando evidenciar as características que poderiam interferir no montante capturado:

1) por dias de captura e;

2) por número de lances de pesca.

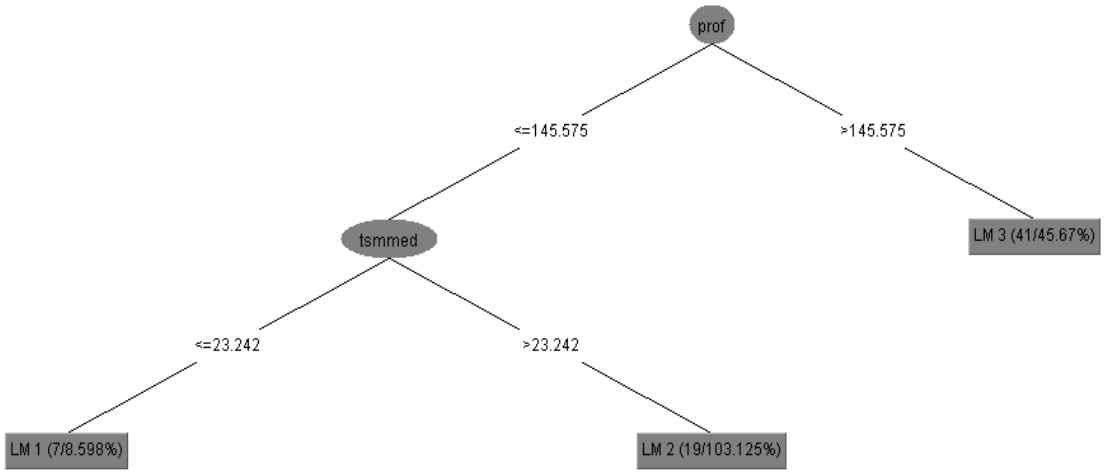

Figura 5: Exemplo de árvore de regressão.

Após a abordagem contínua do atributo alvo, efetuou-se uma abordagem discreta do atributo classe (captura) com o objetivo de evidenciar de forma mais clara as características que podem influenciar no processo de decisão. Dessa forma, podem ser aplicados algoritmos de classificação que relacionam atributos preditivos com um atributo alvo categórico, como por exemplo, algoritmo de árvores de decisão [4].

O atributo alvo foi discretizado em três classes de acordo como conhecimento do especialista: baixas, médias e altas de captura, da seguinte forma:

- 1000 < Captura (kg) < 10000 = baixa;
- 10000 < Captura (kg) < 20000 = média;
- 20000 < Captura (kg) = alta;

Dessa maneira, utilizando o algoritmo de classificação por árvores de decisão C4.5, implementado no Weka como algoritmo J48 [5], com o agrupamento dos dados por dias de pesca e por lances de pesca, foi possível mapear relações diretas entre a classe alvo (captura) e os atributos preditivos considerados (Figura 6).

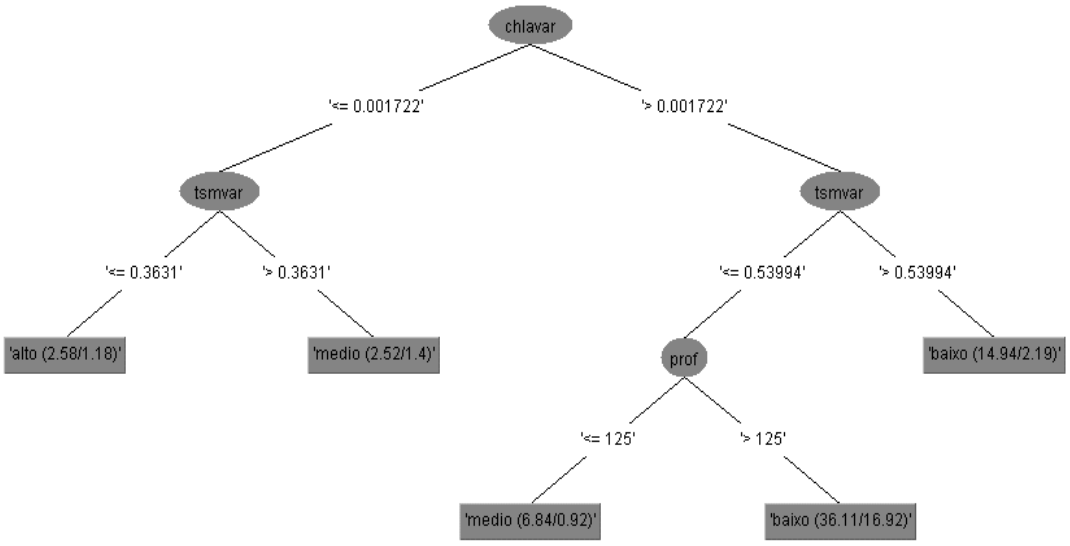

Figura 6: Árvore de decisão obtida utilizando árvores de decisão (algoritmo C4.5), para dias de captura.

3. RESULTADOS E DISCUSSÃO

Correlacionando os parâmetros ambientais e de captura, foi possível perceber a relação entre as variáveis ambientais de TSM e Colorofila a, que haviam sido explorados em outros trabalhos [1] assim como mapear fatores que possivelmente influenciam na produtividade da pescaria avaliada.

Na abordagem por dias de pesca, conforme é mostrada na Figura 7, a alta produtividade desta espécie de tunídeo pode estar associada a profundidade local e a variabilidade térmica da superfície do mar, que favorece aspectos de produtividade trófica, fatores que são de conhecimento científico e também empírico, que leva os mestres de pesca a setores onde mais frequentemente são obtidos maiores rendimentos.

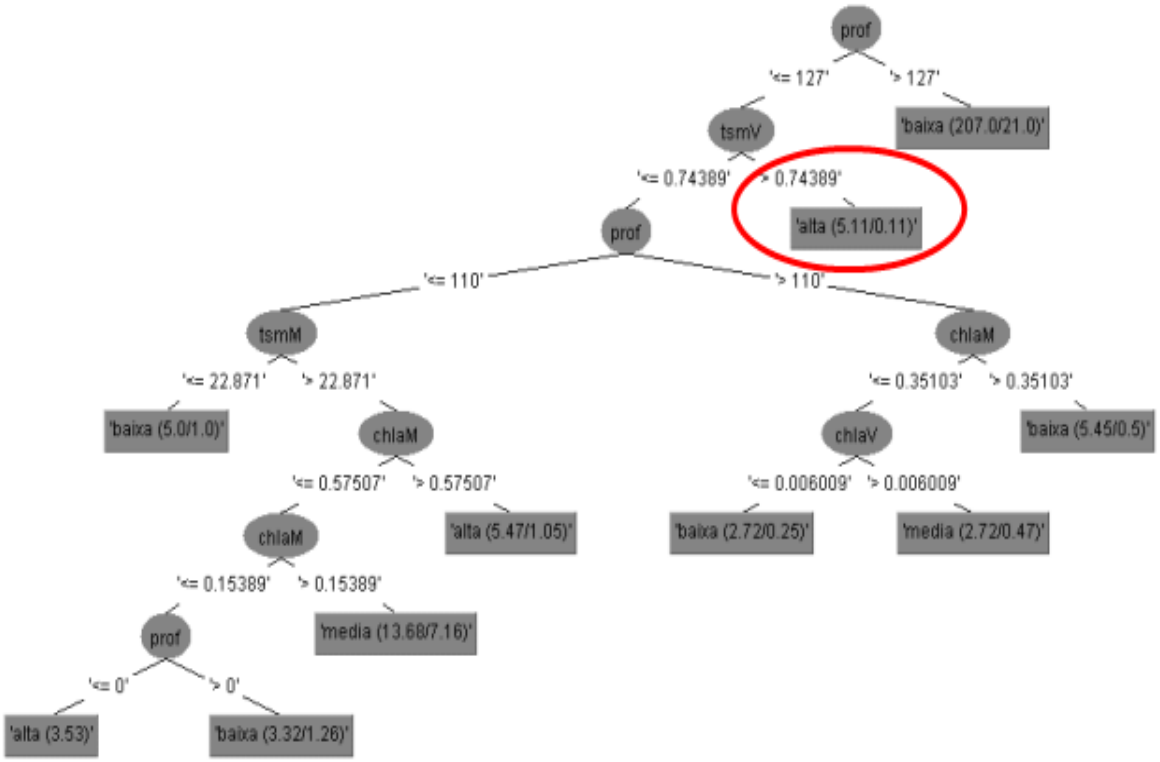

Figura 7: Exemplo de árvore de regressão

Também é possível evidenciar (Figura 8) que a produtividade mostrou-se alta em condições de temperatura e clorofila-a elevadas, o que novamente mostra-se coerente, pois alta concentração de clorofila favorece o desenvolvimento de níveis superiores do plâncton, base da cadeia trófica de pequenos crustáceos e peixes pelágicos, que por sua vez constituem a base alimentar da espécie avaliada. Quanto às temperaturas mais elevadas, há necessidade de maiores informações para estabelecer associações.

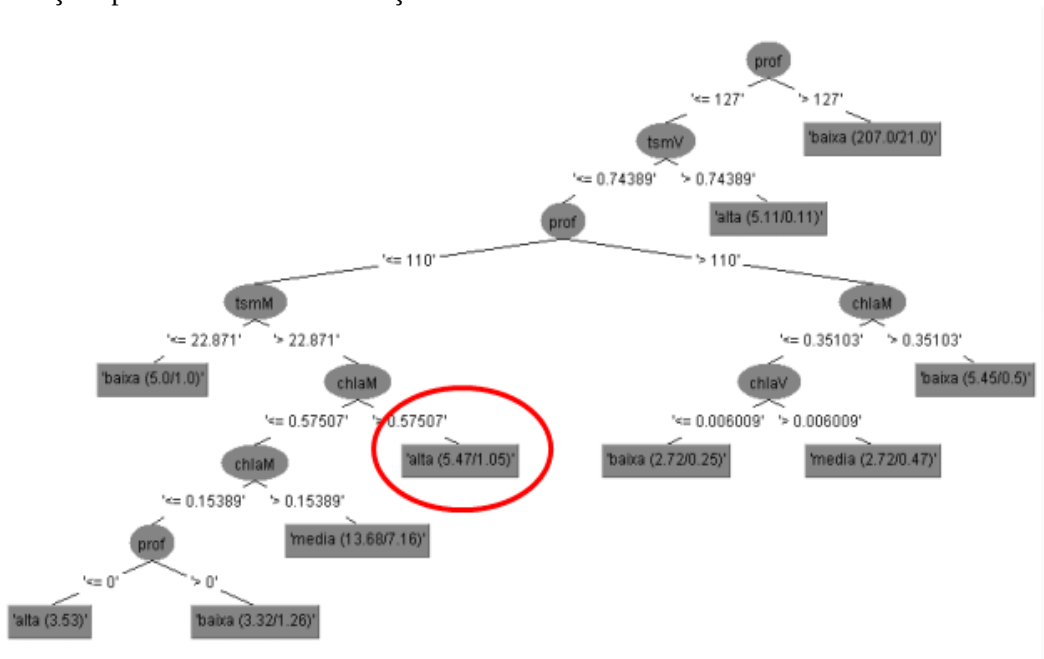

Figura 8: Árvore de decisão para produtividade do atum

Focando no agrupamento por números de lances, as características observadas são mantidas. Porém, novas relações são observadas, como a relação direta da baixa produtividade relacionada a profundidade do lance de pesca, conforme mostrado na Figura 9. Fato este devido ao numero de instâncias consideradas. Na abordagem por dias, como ocorria mais de um lance de pesca por dia, os lances foram agrupados, resultando em dados médios e consequentemente diminuindo o número de instâncias em aproximadamente 50%.

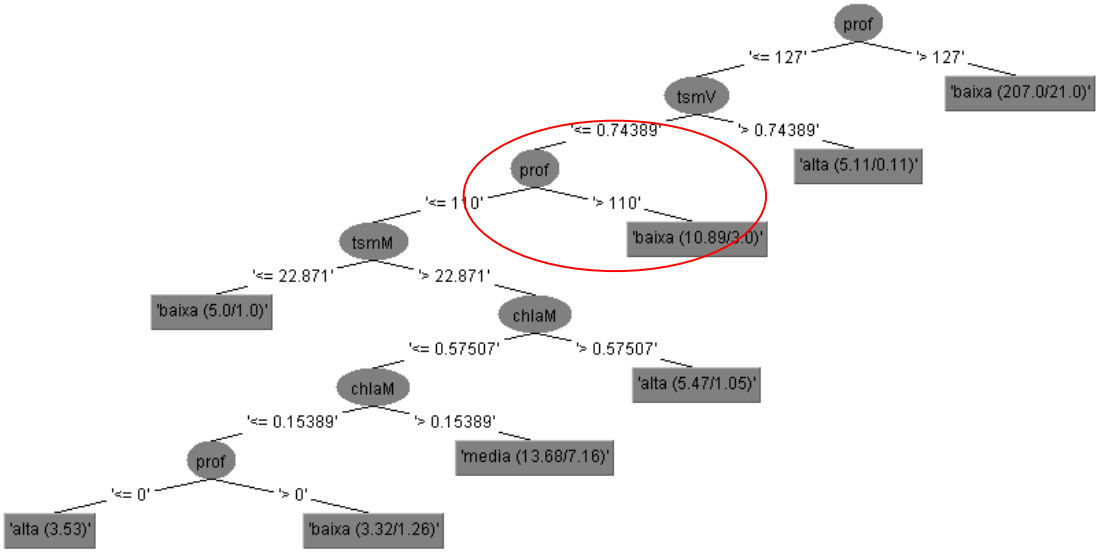

Figura 9: Árvore de decisão por número de lances, obtida com algoritmo de árvores de decisão C4.5 implementação J48 no Weka.

Nas Figuras 10 e 11, podem-se observar as árvores de regressão sem poda e com poda que são obtidas respectivamente sobre a mesma base. Percebe-se que as relações intrínsecas entre os atributos mantêm-se, porém a árvore torna-se mais complexa devido a grande faixa de valores que o atributo classe pode assumir.

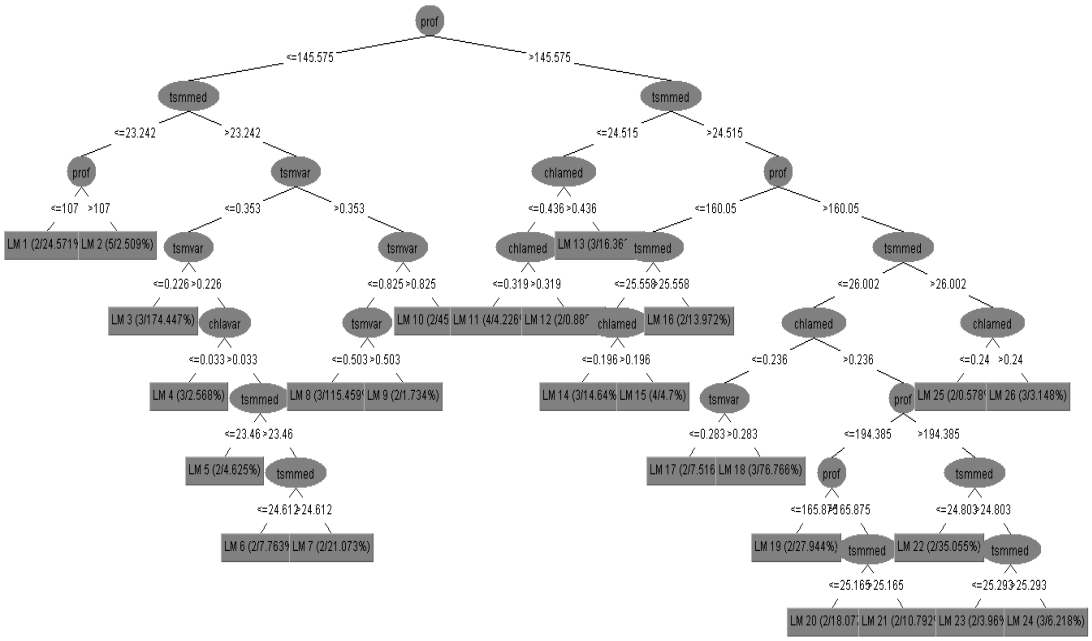

Figura 10: Árvore de regressão de produtividade do atum Bonito listrado sem poda

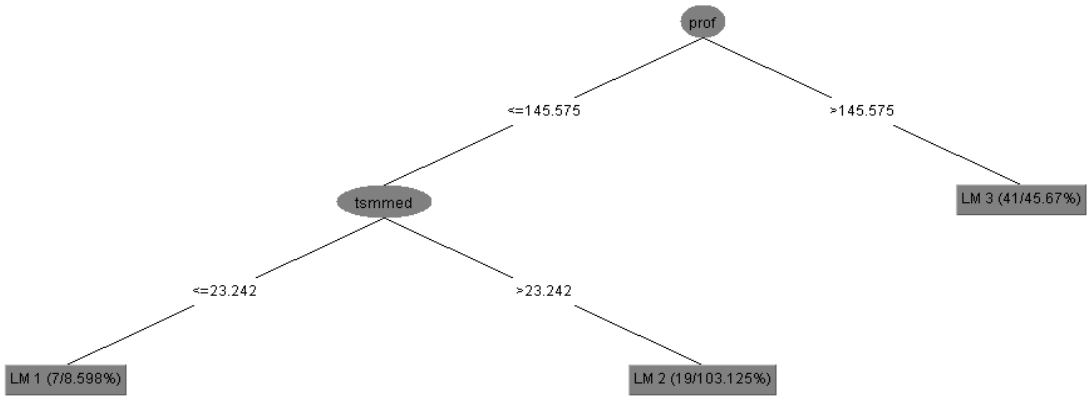

Figura 11: Arvore de regressão de produtividade do Atum com poda

Na Tabela 2, são mostrados dois exemplos de modelos lineares obtidos nas folhas da árvore de regressão. Como é possível observar os modelos, tendem a relacionar os atributos de forma a obter o atributo-alvo captura.

Tabela 2. Exemplo de modelos lineares para árvore de regressão gerada

LM num: 27	LM num: 21
captura = 3923.2968 * lat + 26624.1148 * tsmvar - 202.9609 * chlamed - 122750.8715	captura = 3923.2968 * lat + 845.2408 * tsmmed + 27379.4636 * tsmvar - 108.5799 * prof - 120603.9304

Em relação à qualidade dos modelos de árvore de decisão obtidos través do algoritmo j48, utilizando-se validação cruzada de 10 partições foi obtido o seguinte resultado (Tabela 3).

Tabela 3. Tabela de Validação cruzada gerada pelo software Weka.

Validação Cruzada Estratificada		
Instâncias Classificadas Corretamente	222	87.4%
Instâncias Classificadas Incorretamente	32	12.6%
Estatística Kappa	0.565	
Erro absoluto médio	0.10	
Erro médio quadrático	0.22	
Total de Instâncias	254	

Foram classificados corretamente aproximadamente 87% das instancias, conforme mostrado na matriz de confusão (Tabela 4).

Tabela 4. Matriz de confusão

A	B	C	Classificado como:
197	1	7	A = baixa
21	10	0	B = média
2	0	15	C = Alta

Todos os modelos e avaliações apresentados neste trabalho foram obtidos utilizando-se o software de mineração de dados Weka2.

[2] http://www.cs.waikato.ac.nz/ml/weka/documentation.html

4. CONCLUSÃO

Os resultados obtidos ainda serão analisados por especialistas de modo a explorar novos aprimoramentos e melhorar a confiabilidade e o grau de informação gerado. Deve ficar claro que há necessidade de uma maior entrada de dados nos bancos, exigência esta associada a enorme complexidade dos processos oceanográficos envolvidos, frente a capacidade de amostrá-los. Além disso, os dados brutos, provenientes dos diferentes sensores, contêm erros intrínsecos aos sistemas, seja pela presença de nuvens, no caso dos satélites, seja pela possível inconsistência nos dados de captura de pescado.

Concluindo, neste trabalho foi possível exemplificar e evidenciar a complexidade de um processo de extração de conhecimento, através da abordagem prática de todas as etapas, desde a obtenção dos dados, pré-processamento, mineração dos dados e análise dos resultados, observando que diferentes abordagens sobre o mesmo problema podem evidenciar características distintas.

5. AGRADECIMENTOS

Agradecimentos a equipe do Laboratório de Tecnologia Pesqueira e Hidroacústica (TECPESQ) da Universidade Federal do Rio grande – FURG.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. Andrade HA. 2003. The relationship between the skipjack tuna (Katsuwonus pelamis) fishery and seasonal temperature variability in the south-western Atlantic. Fish. Oceanogr. 12(1): 10-18.
2. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. 1996. In Fayyad UM, Piatetsky-Shapiro G, Smyth P, thurusamy R, editors, Advances in Knowledge Discovery and Data Mining, p.: 1-34. AAAI Press/MIT Press, Menlo Park, CA.
3. Introduction to Data Mining - First Edition, by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, ISBN-13:978-0321321367.
4. Quinlan 1992, Ross J. Quinlan – Learning with Continuous Classes. 5th Australian Joint Conference on Artificial Inteligence, Singapore. 1992.
5. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. "The WEKA Data Mining Software: An Update " SIGKDD Explorations. 2009 July; 11(1):10-18.
6. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005.