Michel de Almeida Silva

O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica

Brasil

2014

Michel de Almeida Silva

O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como requisito parcial para obtenção do título de Mestre em Modelagem Computacional de Sistemas.

Universidade Federal do Tocantins – UFT

Programa de Pós-Graduação Modelagem Computacional de Sistemas

Orientadores
David Nadler Prata e
Patrick Letouze Moreira

Brasil
2014

Michel de Almeida Silva
O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica/ Michel de Almeida Silva. – Brasil, 2014-
83 p. : il. (algumas color.) ; 30 cm.

Orientadores
David Nadler Prata e
Patrick Letouze Moreira

Dissertação de Mestrado – Universidade Federal do Tocantins – UFT
Programa de Pós-Graduação Modelagem Computacional de Sistemas, 2014.

1. Mineração de Dados. 2. Pré-Processamento. I. David Nadler Prata e Patrick Letouze Moreira. II. Universidade Federal do Tocantins. III. Programa de Pós-Graduação em Modelagem Computacional de Sistemas. IV. O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica

CDU 02:141:005.7

Michel de Almeida Silva

O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como requisito parcial para obtenção do título de Mestre em Modelagem Computacional de Sistemas.

Trabalho aprovado. Brasil, 23 de julho de 2014:

Prof. Dr. David Nadler Prata
PPG MCS - UFT
Orientador

Prof. Dr. Patrick Letouze Moreira
PPG MCS - UFT
Orientador

Prof. Dr. George França dos Santos
PPG MCS - UFT
Membro Interno

Prof. Dr. George L. Ribeiro de Brito
PPG MCS - UFT
Membro Interno

Prof. Dra. Selma R. Martins Oliveira
PPG MCS - UFT
Membro Interno

Prof. Dr. Maxwell D. B. de Melo
Engenharia Elétrica - UFT
Membro Externo

Prof. Dr. Ivan Ney Alvizuri Romani
Engenharia Elétrica - UFT
Membro Externo

Brasil

2014

Dedico esse trabalho aos meus pais, Antônio e Lúcia. Sei que não é o suficiente, mas espero que de alguma forma com esse gesto eu possa retribuir todo suporte, carinho e paciência dispensados comigo.

Agradecimentos

Agradeço a minha amada esposa, Larissa, que alegremente concordou em casar com um mestrando, mesmo sabendo que eu não poderia estar sempre presente durante o curso. O seu incentivo foi fundamental durante os dias cansativos e as vezes frustrantes que teimavam em minar minha força de vontade.

Aos professores Dr. David Nadler Prata e Dr. Patrick Leutoze Moreira pela oportunidade de desenvolver este trabalho, suas sábias palavras foram o norte que me guiaram ao longo da jornada. São excelentes profissionais, que muito fazem pelo desenvolvimento da pesquisa no Tocantins alcançando patamares cada vez maiores. Tive muita sorte e foi um prazer imenso tê-los como meus orientadores.

A minha parceira de laboratório Daniela Trevisan pelo companheirismo nas horas e horas de estudo, desenvolvimento e testes. Esse trabalho também é seu. Futura mestre, muito obrigado.

Aos queridos professores do Programa de Pós-Graduação em Modelagem Computacional de Sistemas e em especial a professora Dr. Elineide Marques e professor Dr. Marcelo Lisboa que muito contribuíram com o trabalho, sempre com muita paciência e dedicação.

Aos amigos e familiares que me apoiaram e entenderam minhas ausências durante o curso, em especial aos meus amigos Fred e Brad pela ajuda com os meus inúmeros artigos a serem lidos, vocês são demais.

Aos colegas da DISIN Labs pelo incentivo e colaboração com debates sobre alguns temas que renderam muitas ideias para o projeto e ao M.e Álvaro Prestes, que também faz parte dessa equipe, por toda ajuda que recebi antes e durante o mestrado.

Agradeço ainda a Assembleia Legislativa do Estado do Tocantins e ao Diretor de Área Informática, Carlos Leão, por me permitirem o aprimoramento dos estudos e investirem no servidor da casa.

"Essentially, all models are wrong, but some are useful." - Box, George E. P.

Resumo

Neste trabalho, demonstrou-se que o pré-processamento é um requisito para a Mineração de Dados em conjuntos de dados que apresentam valores errados, ruídos, dados ausentes, inconsistências, redundância e estejam distribuídos em múltiplas fontes heterogêneas.Para tanto, executou-se um estudo sobre um conjunto de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal. Neste caso, utilizou-se o algoritmo *Apriori* sobre os dados pré- processados, o que permitiu a geração de regras de associação que proporcionaram a descoberta de conhecimento singular para o entendimento do processo de desova dos peixes na bacia do rio Tocantins.O segundo estudo realizado utilizou bases de dados do Governo Federal do Brasil, mais especificamente os dados do Programa de Assistência Social Cadastro Único (CadÚnico), formada por dados socioeconômicos de famílias brasileiras de baixa renda, para detectar fatores que são determinísticos para a prática do trabalho infantil e suas relações com indicadores financeiros, educacionais, culturais e sociais, gerando informações que antes não estavam explícitas aos responsáveis pela base de dados. Para a tarefa de classificação dos dados o algoritmo de Mineração de Dados *J48* foi aplicado sobre os dados.Em ambos estudos ficou demonstrada a utilização das técnicas de pré- processamento como limpeza, integração, redução e transformação de dados como suporte para uma Mineração de Dados eficaz na exploração dos dados, descobrindo padrões interessantes do ponto de vista de cada domínio e sugerindo que métodos algorítmicos podem ser empregados em conjunto com métodos estatísticos antes mesmo da geração das hipóteses, contribuindo para esse processo e levando a uma descoberta de conhecimento de qualidade.

Palavras-chaves: mineração de dados. pré-processamento. apriori. ictioplâncton. j48. trabalho infantil. análise exploratória

Abstract

In this study, it was demonstrated that the preprocessing is a requirement for the Data Mining in data sets that show wrong values, noise, missing data, inconsistencies, redundancies and are distributed in multiple heterogeneous sources. To this end, performed a study on a data set of Ichthyoplankton a freshwater reservoir of the Legal Amazon. In this case, we used the *Apriori* algorithm about pre-processed data, allowing association rules generating that provided the discovery of singular knowledge for understanding the process of spawning fish in the Tocantins River basin. The second study used Federal Government databases of Brazil, more specifically the data from the Social Assistance Program Unique Registry (CadÚnico), formed by socioeconomic data families low-income brazilian, to detect factors that are deterministic to the practice of child labour and their relationships with financial indicators, Educational, cultural and social, generating information that before were not express those responsible for the database. For the task of classification the data *J48* Data Mining algorithm was applied over the data. In both studies was demonstrated using the preprocessing techniques such as cleaning, integration, reduction and transforming of data as support for an effective Data Mining on data exploration, discovering interesting patterns from the perspective of each domain and suggesting that algorithmic methods can be employed in conjunction with statistical methods even before the generation of hypotheses contributing to this process, leading to a discovery of quality knowledge.

Key-words: data mining. pre-processing. apriori. ichthyoplankton. j48. child labor. exploratory analysis.

Lista de ilustrações

Figura 1 – Visão geral das fases do KDD. 21
Figura 2 – Taxonomia dos métodos de Mineração de Dados segundo Maimon e Rokach (2010). 22
Figura 3 – Exemplo de árvore de decisão criado sobre as condições de jogar ou não um jogo qualquer (WITTEN; FRANK; HALL, 2011). 27
Figura 4 – A Investigação Interdisciplinar modelo de Gestão de Projeto. (Fonte: Traduzida de (LETOUZE, 2011)) 37
Figura 5 – Diagrama do processo utilizado no presente trabalho. 43

Lista de abreviaturas e siglas

API	*Application Programming Interface* (Interface de Programação de Aplicativos)
ARFF	*Attribute-Relation File Format* (Formato de Arquivo Atributo-Relação)
CadÚnico	Programa de Assistência Social Cadastro Único
CSV	*Comma-Separated Values* (Valores Separados por Vírgula)
EODIS	*Earth Observing System Data and Information System* (Sistema de Informação e de Dados de Observação da Terra)
GNU	Acrônimo recursivo de para *GNU is Not Unix* (GNU Não é Unix)
IBGE	Instituto Brasileiro de Geografia e Estatística
ICDM	*IEEE International Conference on Data Mining series* (Conferência Internacional sobre séries de Mineração de Dados da IEEE)
IEEE	*Institute of Electrical and Electronics Engineers* (Instituto de Engenheiros Eletricistas e Eletrônicos)
ILO	*International Labour Organisation* (Organização Mundial do Trabalho Infantil)
IRPM	*Interdisciplinary Research Project Management* (Gerência de Projetos de Pesquisa Interdisciplinar)
JAR	*Java ARchive* (ARquivo Java)
JDBC	*Java Database Connectivity* (Conectividade de Banco de Dados Java)
JDK	*Java Development Kit* (Kit de Desenvolvimento Java)
KDD	*Knowledge Discovery in Database* (Descoberta do Conhecimento em Banco de Dados)
MDS	Ministério do Desenvolvimento Social e Combate à Fome
NASA	*National Aeronautics & Space Administration* (Administração Nacional da Aeronáutica e do Espaço)
PBF	Programa Bolsa Família

PBL	*Problem Based Learning* (Aprendizado Baseado em Problemas)
SKICAT	*Sky Image Cataloging and Analysis Tool* (Ferramenta de Análise e Catalogação de Imagem do Céu)
SQL	*Structured Query Language* (Linguagem de Consulta Estruturada)
UNICEF	Fundo das Nações Unidas para a Infância
Weka	*Waikato Environment for Knowledge Analysis* (Ambiente Waikato para Análise de Conhecimento)
XLS	*MS Excel file extension* (Extensão de arquivo Microsoft Excel)

Sumário

1 INTRODUÇÃO 15
1.1 Visão Geral do Problema 16
1.2 Justificativa 17
1.3 Objetivo Geral 18
1.3.1 Objetivos Específicos 18
1.4 Organização da Dissertação 18

2 FUNDAMENTAÇÃO TEÓRICA E ESTADO-DA-ARTE 20
2.1 Fundamentação Teórica 20
2.1.1 Mineração de Dados 20
2.1.2 Pré-Processamento 23
2.1.3 Seleção de atributos baseada em correlação 24
2.1.4 Regras de associação 24
2.1.5 Classificação com árvores de decisão 26
2.2 Estado-da-Arte 29
2.3 Trabalhos Importantes para Área 34

3 METODOLOGIA 36
3.1 IRPM 36
3.1.1 Aplicação do IRPM 38
3.2 Materiais 39
3.2.1 Equipamentos 40
3.2.2 Softwares 40
3.2.3 Sistema Operacional 41
3.3 Processos 41

4 APLICAÇÃO DO ALGORITMO APRIORI PARA UMA BASE DE DADOS DE ICTIOPLÂNCTON EM UM RESERVATÓRIO DE ÁGUA DOCE DA AMAZÔNIA LEGAL 44

5 EXPLORANDO DADOS SOCIAIS PARA ENTENDER O TRABALHO INFANTIL 57

6 MINERAÇÃO DE DADOS COMO FERRAMENTA DE ANÁLISE EXPLORATÓRIA EM SISTEMAS ECOLÓGICOS 67

7 CONCLUSÕES 79

Referências 81

1 Introdução

O sucesso em Mineração de Dados é medido pela qualidade dos resultados. Pesquisadores envolvidos com esse tipo de modelagem algorítmica buscam resultados mais precisos durante a aplicação de algoritmos utilizados para descoberta do conhecimento induzido diretamente a partir de dados. Então a qualidade dos dados de entrada é um fator determinante na qualidade do conhecimento extraído.

Os bancos de dados do mundo real são altamente suscetíveis a ruídos (erros e exemplos fora do padrão), valores errados, dados ausentes, inconsistências e dados redundantes que muitas vezes estão distribuídos em múltiplas fontes heterogêneas, que são fatores que reduzem a qualidade dos dados e consequentemente reduzem a qualidade dos resultados e a eficiência da Mineração de Dados. Nesse contexto as técnicas de pré-processamento fornecem suporte as pesquisas algorítmicas melhorando a eficiência e facilitando o processo de Mineração de Dados, ao mesmo tempo que permitem ao pesquisador um melhor entendimento da natureza dos dados a serem "minerados".

Em particular, existem duas abordagens na criação de modelos para tirar conclusões a partir de dados, segundo Breiman et al. (2001), a modelagem de dados em que os dados são gerados por um modelo estocástico e a modelagem algorítmica que usa modelos algoritmos e trata o mecanismo de dados como desconhecido.

Nesse contexto a Mineração de Dados é um exemplo de modelagem algorítmica que desde sua concepção na década de 80 tem alcançado enorme sucesso em pesquisas científicas. Ela integra um processo maior, conhecido por Descoberta do Conhecimento (*Knowledge Discovery in Database – KDD*) definida por Fayyad et al. (1996) como um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis. Por outro lado, o KDD segundo Hong, Kuo e Chi (1999) compreende três fases a seguir:

- Pré-Processamento: consiste na preparação dos dados e todas medidas necessárias para iniciar a Mineração de Dados. Por causa das origens diversas é comum que os dados não estejam prontos para que os algoritmos sejam aplicados diretamente. Durante essa preparação o pesquisador tem a oportunidade de obter um melhor entendimento da natureza dos dados a serem processados;

- Mineração de Dados: aplicação de algoritmos próprios para extrair padrões ou regras de uma base de dados;

- Pós-Processamento: os padrões descobertos ou as regras encontradas são traduzidas para uma forma aceitável para o entendimento humano.

Embora existam problemas fundamentais a serem resolvidos como o desenvolvimento de uma teoria unificada (YANG; WU, 2006), os benefícios da mineração foram demonstrados em uma ampla gama de domínios de aplicação. Por exemplo na astronomia analisando dados sobre o planeta Terra recolhidos pela projeto *Earth Observing System Data and Information System* (EOSDIS) da NASA (*National Aeronautics & Space Administration*) com seu 1 petabyte de dados sobre a Terra por ano (NASA, 2014), nas áreas médica e biológica (MOORE; RITCHIE, 2013), utilizando bancos de dados públicos para fornecer *insights* sobre desenvolvimento de drogas ou com temas em biodiversidade com foco em desenvolvimento de códigos de barra para todos organismos em um ecossistema particular.

Quando o conhecimento acerca do domínio é insuficiente para extrair percepções dos dados as análises exploratórias são mais apropriadas e nesse contexto entra a modelagem algorítmica que, de acordo com Breiman et al. (2001), é caracterizada como uma análise de dados que:

a) automaticamente faz predições precisas a partir dos dados;

b) possui capacidade para lidar com um grande número de variáveis preditoras e de identificação das mais importantes;

c) trata o modelo como uma caixa preta onde o pesquisador não tem necessidade de fazer suposições sobre o relacionamento das variáveis preditoras e as de resposta.

A modelagem algorítmica de dados já pré-processados, dessa maneira, pode ser utilizada também para identificar componentes importantes que poderão ser parâmetros em outras análises. Dessa forma o pré-processamento e a Mineração de Dados podem ser combinados com análises estatísticas para extrair soluções ainda melhores do que as geradas apenas com a utilização da modelagem de dados.

1.1 Visão Geral do Problema

Diferentemente dos conjuntos de dados que acompanham os exemplos em livros de Mineração de Dados (WITTEN; FRANK; HALL, 2011) os bancos de dados do mundo real são suscetíveis a:

i) valores errados;

ii) ruídos;

iii) dados ausentes;

iv) inconsistências, devido ao seu enorme tamanho (petabytes ou mais, como no exemplo anterior do projeto da NASA);

v) dados redundantes;

vi) distribuídos em múltiplas fontes heterogêneas.

Há muitos fatores que compõem a qualidade dos dados, incluindo a exatidão, integridade, consistência, pontualidade, credibilidade e facilidade de interpretação (HAN; KAMBER; PEI, 2006). Para alcançar resultados de qualidade em Mineração de Dados uma das partes mais importantes em qualquer projeto, e uma das mais demoradas e difíceis, é o pré-processamento, que leva em torno de 70% do tempo dispendido à pesquisa (FERNANDEZ, 2002). Portanto, o pré-processamento é indispensável para aplicação dos algoritmos de Mineração de Dados, tanto para aumentar a qualidade dos resultados e diminuir o tempo destinado a mineração, como para um melhor entendimento da natureza dos dados a serem minerados.

1.2 Justificativa

Em pesquisas científicas eventualmente há alguma coleta de dados e muitas vezes bases de dados são formadas a respeito do tópico pesquisado. Assim sendo, métodos algorítmicos podem ser utilizados como alternativa aos métodos estatísticos tradicionais na modelagem do problema. Algumas vezes, esses métodos podem prover resultados interessantes a partir dessas bases de dados. No entanto, a qualidade desses dados pode requerer um pré-processamento de modo a dar suporte a modelagem algorítmica e melhorar a qualidade dos resultados obtidos.

A análise de dados é realizada com a finalidade de descobrir padrões e extrair informações úteis que geram conhecimento necessário para tomadas de decisões de modo a compor o que vem a ser Mineração de Dados. Para que essa análise seja realizada de maneira eficiente o conjunto de dados deve passar pela etapa de pré-processamento (KANTARDZIC, 2011).

O pré-processamento é muitas vezes a chave para resolver o problema. Pode ser facilmente a diferença entre uma modelagem algorítmica realizada com sucesso e a fracassada, entre predições de valor e suposições inúteis, além de ocupar boa parte do tempo da pesquisa. Assim, se o pré-processamento não for realizado corretamente, pode inviabilizar a informação utilizável.

1.3 Objetivo Geral

Demonstrar que o pré-processamento de dados faz-se necessário como método de suporte a modelagem algorítmica, em particular, a Mineração de Dados para conjuntos de dados de ictioplâncton e bases de dados do Governo Federal do Brasil.

HIPÓTESE: "o pré-processamento de dados é um requisito para a Mineração de Dados em conjuntos de dados que apresentam valores errados, ruídos, dados ausentes, inconsistências, redundância e estejam distribuídos em múltiplas fontes heterogêneas específicamente nas bases de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal e bases de dados do Programa de Assistência Social Cadastro Único (CadÚnico) do Governo Federal".

1.3.1 Objetivos Específicos

- Demonstrar a importância do pré-processamento para melhorar a qualidade dos resultados obtidos na Mineração de Dados;
- Explorar bases de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal;
- Explorar bases de dados do Programa de Assistência Social Cadastro Único (CadÚnico) do Governo Federal;
- Demonstrar a utilidade da Mineração de Dados para realização da análise exploratória;

1.4 Organização da Dissertação

Esta dissertação encontra-se organizada em 7 capítulos, incluindo esta introdução. A seguir, apresenta-se uma breve descrição de cada um dos capítulos subsequentes:

- Capítulo 2: apresenta a Fundamentação Teórica e o Estado-da-Arte que fornecem fundamentação à pesquisa e as contribuições proporcionadas por investigações anteriores em Mineração de Dados e no pré-processamento.
- Capítulo 3: apresenta a Metodologia, onde expõe-se o IRPM como método de pesquisa, lista-se os materiais utilizados e os processos desenvolvidos durante a pesquisa.
- Capítulo 4: apresenta uma versão em português do artigo *Exploring an Ichthyoplankton Database from a Freshwater Reservoir in Legal Amazon* (SILVA et al., 2013) publicado como artigo em *Lecture Notes in Computer Science* 2013 com classificação

Qualis Capes: A1 - Biodiversidade e B1 - Interdisciplinar. O artigo foi publicado com intuito de introduzir o uso de técnicas de Mineração de Dados incluindo o pré-processamento, para análise exploratória em uma base de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal.

- Capítulo 5: apresenta uma versão em português do artigo *Exploring Social Data to Understand Child Labor* (RODRIGUES; PRATA; SILVA, 2013) publicado no *International Journal of Social Science and Humanity* 2013 com classificação Qualis Capes: B1 - Interdisciplinar. O artigo publicado com o uso de técnicas de Mineração de Dados para descobrir padrões válidos entre bases de dados sociais brasileiros, na tentativa de descobrir indicadores que levam ao trabalho infantil.

- Capítulo 6: apresenta o artigo Mineração de Dados como ferramenta de análise exploratória em Sistemas Ecológicos a ser submetido para publicação. O artigo demonstra que técnicas de Mineração de Dados podem ser utilizadas para fins de análise exploratória em pesquisas ecológicas onde existem conjuntos de dados como alternativa a modelagem de dados tradicional.

- Capítulo 7: apresenta as conclusões do trabalho desenvolvido, bem como sugestões de alguns trabalhos que podem ser realizados futuramente.

2 Fundamentação Teórica e Estado-da-Arte

Este capítulo é dividido em duas partes, a primeira apresenta a Fundamentação Teórica da dissertação onde é realizada uma abordagem geral sobre Mineração de Dados, pré-processamento e os algoritmos *CfsSubsetEval* para seleção de atributos baseada em correlação, Apriori para regras de associação e J48 para classificação com árvores de decisão, apresentando uma descrição de cada algoritmo e seu funcionamento dentro do objetivo de cada técnica. A segunda parte apresenta o Estado-da-Arte com a discussão de trabalhos importantes para área.

2.1 Fundamentação Teórica

Nesta seção apresenta-se a fundamentação teórica referentes aos conceitos de Mineração de Dados, pré-processamento, seleção de atributos baseada em correlação, regras de associação, classificação com árvores de decisão.

2.1.1 Mineração de Dados

A quantidade de dados está cada vez maior e deve continuar aumentando com a redução do custo dos equipamentos necessários para armazenamento e com a disponibilização de serviços que fazem isso de forma "online" através de *cloud computing* (computação nas nuvens). Esse crescimento no volume dos dados remete a uma questão que é: como explorar esses dados e extrair informação?

A exploração de dados para extrair padrões e realizar previsões não é algo novo, segundo Witten, Frank e Hall (2011) economistas, estatísticos, analistas e engenheiros de comunicação há muito tempo trabalham com a ideia de que padrões nos dados podem ser procurados automaticamente, identificados, validados e utilizados para fins de previsão. O que é novo é o aumento vertiginoso em oportunidades para encontrar padrões em dados. O crescimento desenfreado de bases de dados nos últimos anos, traz a mineração de dados para a vanguarda das novas tecnologias de negócios.

Apesar de Mineração de Dados ser o termo mais conhecido, ele é na verdade parte de um processo maior denominado Descoberta do Conhecimento, do inglês *Knowledge Discovery in Databases* (KDD), que segundo Fayyad et al. (1996) é o processo não trivial de identificar padrões em dados que sejam válidos, novos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um processo de tomada de decisão.

Fayyad et al. (1996) define a Mineração de Dados como um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados.

A Figura 1 fornece uma visão geral sobre o processo do KDD segundo Hong, Kuo e Chi (1999) e situa o papel da Mineração de Dados no processo como um todo.

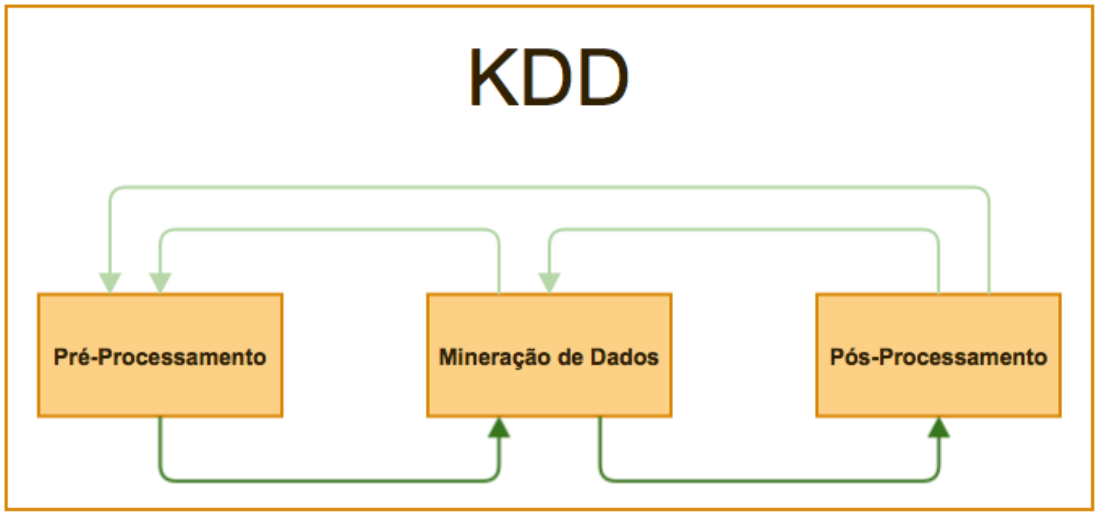

Figura 1 – Visão geral das fases do KDD.

Como uma das fases do processo KDD a Mineração de Dados é geralmente reconhecida pela capacidade de realizar algumas tarefas (LAROSE, 2014), as mais comuns são:

- Descrição (Description): é a tarefa utilizada para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

- Classificação (Classification): uma das tarefas mais comuns, a classificação, visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de "aprender" como classificar um novo registro (aprendizado supervisionado).

- Estimação (Estimation): a estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um nominal. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais.

- Predição (Prediction): a tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo.
- Agrupamento (Clustering): a tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares.
- Associação (Association): a tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: **SE** atributo X **ENTÃO** atributo Y. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras" (*Market Basket*), onde identificamos quais produtos são levados juntos pelos consumidores.

Existem inúmeros métodos para realização dessas tarefas em Mineração de Dados, por vezes o método a ser aplicado depende do tipo de dado de entrada e o tipo de informação desejada como saída. Os métodos se dividem entre métodos de verificação e métodos de descoberta. Os de descoberta se dividem ainda em métodos de predição e métodos de descrição. Esta divisão está ilustrada na Figura 2.

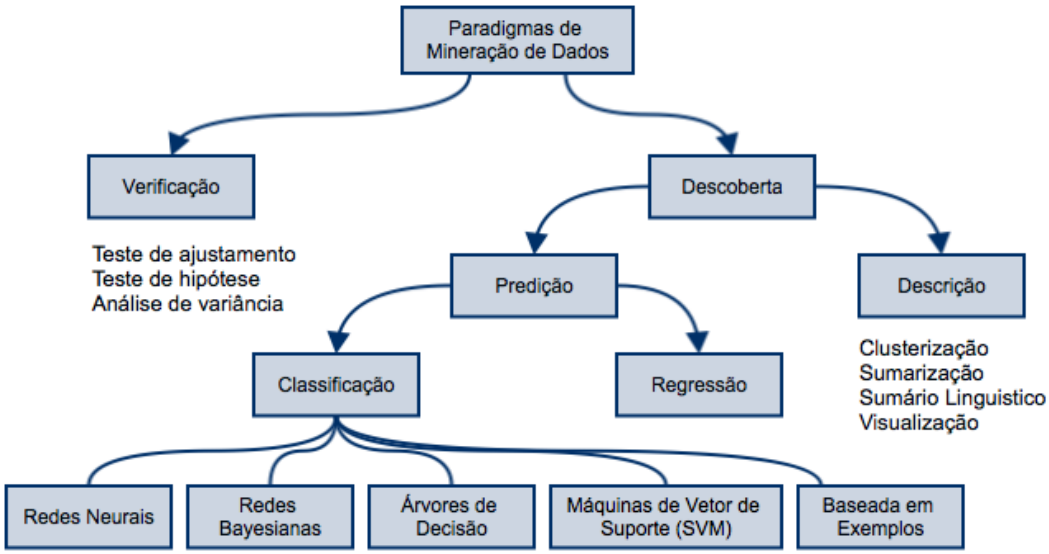

Figura 2 – Taxonomia dos métodos de Mineração de Dados segundo Maimon e Rokach (2010).

Todos esses métodos podem ser utilizados sozinhos ou em conjunto, mas não existe uma regra que indique qual é o método correto para um determinado problema. A escolha do método está ligada a tarefa da Mineração de Dados e ao tipo de variável.

2.1.2 Pré-Processamento

Bancos de dados reais de hoje são altamente suscetíveis a ruídos, ausência de dados e inconsistentes devido a seu tamanho, geralmente grande (muitas vezes de vários terabytes ou mais), e sua provável origem de múltiplas fontes heterogêneas. Consequetemente, é possível que a qualidade desses dados esteja comprometida, e dados de baixa qualidade irão levar a resultados de Mineração de Dados de baixa qualidade. Portanto, necessita-se de uma fase de uma fase anterior, que consiste na preparação dos dados para que possam ser aplicados os algoritmos de mineração, essa fase é denominada pré-processamento (HAN; KAMBER; PEI, 2006). As principais etapas envolvidas no pré-processamento são:

- Limpeza dos dados: essa etapa de limpeza dos dados visa eliminar os problemas como registros incompletos, valores errados e dados inconsistentes, de modo que eles não influenciem no resultado dos algoritmos usados. As técnicas usadas nesta etapa vão desde a remoção do registro com problemas, passando pela atribuição de valores padrões, até a aplicação de técnicas de agrupamento para auxiliar na descoberta dos melhores valores.

- Integração dos dados: comumente os dados são oriundos de fontes heterogêneas tais como banco de dados, arquivos textos, planilhas, *data warehouses*, vídeos, imagens, etc. Surge então, a necessidade da integração destes dados de forma a termos um repositório único e consistente. Para isto, é necessária uma análise aprofundada dos dados observando redundâncias, dependências entre as variáveis e valores conflitantes, por exemplo categorias diferentes para os mesmos valores, chaves divergentes, regras diferentes para os mesmos dados, entre outros.

- Redução dos dados: o volume de dados usado na mineração costuma ser alto, em alguns casos este volume é tão grande que torna o processo de análise dos dados e da própria mineração impraticável. Assim, as técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, porém sem perder a representatividade dos dados originais. Isto permite que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado. As estratégias adotadas nesta etapa são a criação de estruturas otimizadas para os dados (cubos de dados), a seleção de um subconjunto dos atributos, a redução da dimensionalidade e a discretização.

- Transformação dos dados: alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores nominais, e nestes casos é necessário transformar os valores numéricos em nominais ou os nominais em valores numéricos. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta

etapa são a suavização, o agrupamento, a generalização, a normalização e a criação de novos atributos a partir de outros já preexistente.

2.1.3 Seleção de atributos baseada em correlação

A seleção de atributos baseada em correlação avalia o valor de um subconjunto de atributos considerando a capacidade preditiva individual de cada recurso, juntamente com o grau de redundância entre eles. Os subconjuntos de características altamente correlacionados com a classe e com baixa intercorrelação são preferencialmente selecionados (HALL et al., 2009). O autor Hall (1999) deixa claro que correlação é um termo usado no seu sentido geral. Não se destina para referir-se especificamente a correlação linear clássica. O autor o usa para se referir a um grau de dependência ou previsibilidade de uma variável com a outra.

O *CfsSubsetEval* é o algoritmo que realiza a seleção de atributos baseada em correlação por uma filtragem simples que classifica subconjuntos de atributos de acordo com uma correlação baseada em função de avaliação heurística. A função de avaliação busca subconjuntos quecontêm características que são altamente correlacionados com uma classe e não correlacionadas com outra. Características irrelevantes são portanto ignoradas porque têm baixa correlação.

Para avaliação de atributos é utilizada uma heurística denominada Mérito, formalizada pela equação (2.1) (GHISELLI, 1964).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.1}$$

onde $Merit_s$ é a heurística "Mérito" de um subconjunto S contendo k atributos, $\overline{r_{cf}}$ é a correlação média do atributo classe ($f \in S$), e $\overline{r_{ff}}$ é a correlação média entre os atributos.

O numerador da equação (2.1) pode ser pensado como uma indicação de como um conjunto de atributos é preditivo da classe e o denominador de quanta redundância há entre os atributos. A fórmula final de Mérito utiliza a correlação de Pearson entre uma variável composta (soma ou média) e uma variável de destino (a classe em questão) (HALL, 2000).

2.1.4 Regras de associação

Os algoritmos capazes de encontrar relacionamentos entre os dados são chamados de algoritmos de regras de associação e trabalham com a extração de conjuntos de atributos frequentes inseridos em um conjunto maior. Esses algoritmos variam bastante em relação a geração de subconjuntos e em como os conjuntos de atributos escolhidos são suportados durante a geração das regras de associação.

A regra de associação tem o formato A $\longrightarrow$ B, onde A é chamado de antecedente, e B é chamado de consequente. A e B são conjuntos de itens ou transações, e a regra pode ser lida como: o atributo A frequentemente implica no atributo B (AGRAWAL; IMIELINSKI; SWAMI, 1993). Para avaliar as regras geradas são utilizadas algumas medidas de interesse, as mais utilizadas são suporte e confiança também estabelecidas no trabalho de Agrawal, Imielinski e Swami (1993). Já os autores Geng e Hamilton (2006) realizaram uma pesquisa envolvendo outras métricas e sugeriram estratégias para selecionar medidas adequadas para determinados domínios e exigências. As seguintes medidas são utilizadas para mensurar as regras geradas:

- Suporte (support): $P(AB)$. O suporte de uma regra é definido como sendo a fração de itens I que satisfazem o conjunto A e B da regra. Se o suporte não é grande o suficiente, isso significa que a regra não é digna de consideração ou que é simplesmente preterida e pode ser considerada mais tarde;
- Confiança (confidence): $P(A/B)$. É uma medida da força de suporte as regras e corresponde a significância estatística. A probabilidade de encontrar B da regra nas transações sobre a condição que essas transações também contenham A;
- Interesse (lift): $P(B|A)/P(B)$ ou $P(AB)/P(A) * P(B)$. Utilizada para encontrar dependências, ela indica o quanto mais frequente torna-se B quando A ocorre.

Encontrar conjuntos de itens frequentes, com frequência maior ou igual à especificada pelo usuário como sendo o suporte mínimo, não é trivial, devido à explosão combinatória ocorrida ao gerar os subconjuntos de itens. Mas, uma vez que os conjuntos de itens frequentes são obtidos, é muito simples gerar regras de associação com confiança maior ou igual à especificada pelo usuário como sendo o valor mínimo (WU et al., 2008).

O algoritmo *Apriori* foi introduzido em (AGRAWAL; SRIKANT et al., 1994) como uma maneira de gerar regras de associação de dados.

Basicamente o algoritmo é composto por duas etapas:

(1) Encontrar todos os conjuntos de itens frequentes, ou seja, com valor de suporte superior ou igual ao suporte mínimo especificado pelo usuário. Em termos de custo computacional, esta é a etapa mais custosa;

(2) Usar os conjuntos de itens frequentes para induzir as regras de associação, com valores de suporte e confiança superior ou igual ao suporte e confiança mínimos especificados pelo usuário.

Empregando busca em profundidade o algoritmo é capaz de gerar conjuntos de itens candidatos (reconhecidos como o padrão) com k elementos a partir de conjuntos de

itens de $k \neq 1$ elementos. A varredura só termina no último elemento da base de dados e padrões não frequentes são descartados, o algoritmo 1 demonstra o funcionamento do *Apriori*.

Algoritmo 1: Pseudo código do algoritmo *Apriori*

```
Input: (T,minSupport)
// T é o conjunto de dados e minSupport é o suporte mínimo
begin
    L1={itens frequentes};
    for (k=2; L_{k-1} != ∅; k++ ) do
        C_k = candidatos gerados a partir de L_{k-1};
        // é o produto cartesiano L_{k-1} x L_{k-1} e elimina
        // qualquer conjunto de itens de tamanho k − 1 que não seja frequente
        foreach transação t ∈ T do
            #incrementa o contador de todos candidatos em C_k que estão ⊂ em t;
            L_k = candidatos em C_k com minSupport;
        end
        // fim do for each
    end
    // fim do for
end
return U_k L_k;
```

Como é comum na mineração de regra de associação, dado um conjunto de conjuntos de itens (por exemplo, das operações de varejo de moda, cada listagem individual itens comprados), o algoritmo tenta encontrar subconjuntos, que são comuns a pelo menos um número mínimo C dos conjuntos de itens. O *Apriori* usa uma abordagem "bottom-up", onde subconjuntos frequentes são estendidos um item por vez (uma etapa conhecida como geração de candidatos), e grupos de candidatos são testados contra os dados. O algoritmo termina quando não mais se encontram as estensões de sucesso.

2.1.5 Classificação com árvores de decisão

O método de classificação por Árvore de Decisão, funciona como um fluxograma em forma de árvore (CAMILO; SILVA, 2009), onde cada nó (não-folha) indica um teste feito sobre um valor. As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe a qual o registro pertence. Após a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore (mediante os testes nos nós não-folhas) começando no nó raiz até chegar a uma folha, como mostra o exemplo da Figura 3:

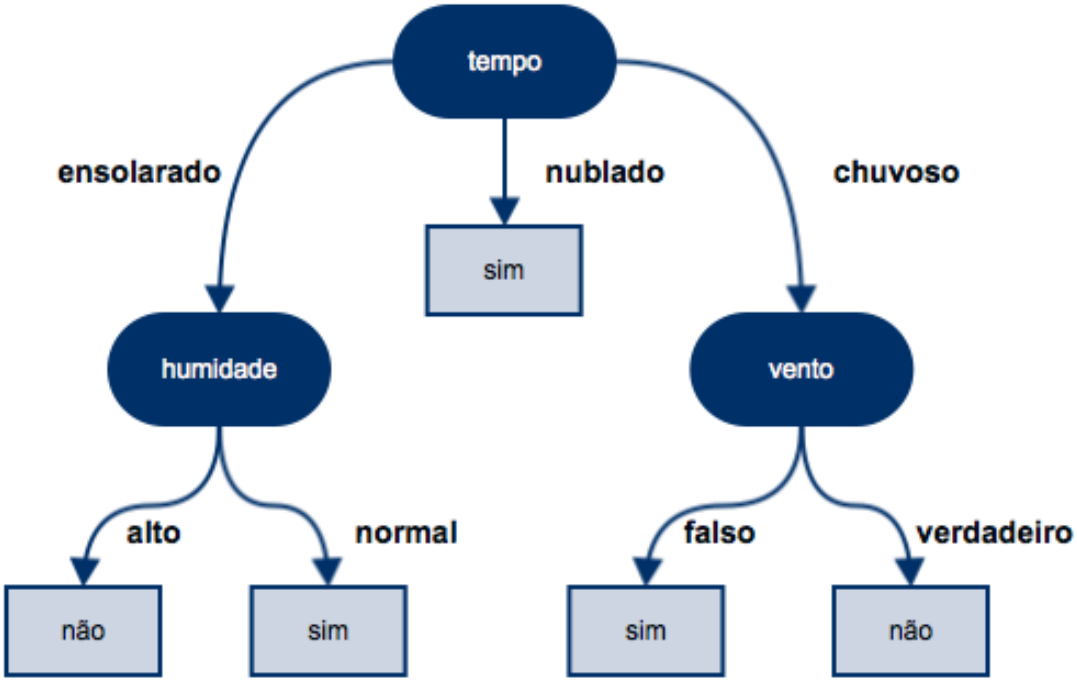

Figura 3 – Exemplo de árvore de decisão criado sobre as condições de jogar ou não um jogo qualquer (WITTEN; FRANK; HALL, 2011).

Com base na árvore de decisão do exemplo da Figura 3 pode-se decidir quando jogar ou não um jogo de acordo com as condições do tempo.

Pela estrutura que formam, as árvores de decisões podem ser convertidas em Regras de Classificação. O sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade.

O algoritmo *J48* (WITTEN; FRANK; HALL, 2011) é uma implementação do algoritmo de árvore de decisão *C4.5* proposto por Quinlan (1993) para o Weka. O algoritmo *J48* constrói a árvore de cima para baixo (*top-down*) com o objetivo de escolher sempre o melhor atributo para cada nó de decisão da árvore. É um processo recursivo que após ter escolhido um atributo para um nó, começando pela raiz, aplica o mesmo algoritmo aos descendentes desse nó, até que os critérios de parada sejam verificados.

A escolha do atributo de partição é realizada pelo valor do ganho de informação. O Ganho de Informação é uma medida estatística que está no cerne da construção de árvores de decisão no algoritmo *J48*. Essa medida consiste em:

Se tivermos um conjunto de vários exemplos S, e um conjunto de n classes $C = C_1, C_2, ..., C_n$, sendo p_i a probabilidade da classe C_i em S, então a entropia do conjunto S, é a homogeneidade deste, traduzida na seguinte igualdade:

$$Entropia = -\sum_{i=1}^{c} p_m \log_2 p_i \tag{2.2}$$

A entropia é uma medida aplicável à partição de um espaço de probabilidade, medindo quanto esse espaço é homogéneo, ou por outro lado, quanto maior a entropia maior a desordem. A entropia atinge o seu valor máximo, igual a $\log_2 n$, quanto $p_1 = p_2 = ... = p_n = 1/n$, expressando precisamente a existência de uma máximo de heterogeneidade. Pelo contrário a homogeneidade máxima corresponderia a $p_1 = p_2 = ... = p_n = 0$ e $p_i = 1$.

De outro modo, pretende-se saber qual o ganho de informação do atributo A, que é dado pela seguinte fórmula:

$$Ganho(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.3}$$

em que, valores (A) é o conjunto de todos os valores possíveis para o atributo A, e $|S_v|$ é o sub-conjunto de S para o qual o atributo A tem valor v:

$$S_v = \{s \in S | A = s\} \tag{2.4}$$

Desta forma, o Ganho de Informação, mede a eficácia de um atributo em classificar os dados de treino, a escolha do atributo mais eficaz – que mais reduz a entropia – faz com que a tendência seja a de gerar árvores, que são, em geral, menos profundas com menos nós e ramificações.

Em suma, o algoritmo *J48* realiza uma busca gulosa (*greedy search*) no espaço das árvores de decisão consistente com os dados, guiada pelo ganho de informação e feita segundo a estratégia "subir a colina" (*hill-climbing*), o algoritmo 2 detalha o funcionamento

do *J48*.

Algoritmo 2: Pseudo código do algoritmo *J48*

```
Input: um conjunto de dados D
begin
    Árvore={};
    if D é "puro" OU existe outro critério de parada then
        encerrar;
    foreach atributo a ∈ D do
        Calcular ganho de informação;
    end
    a_melhor = Melhor atributo de acordo com o calculo do ganho de informação;
    Árvore = Cria um nó baseado no a_melhor;
    D_v = Divide o subconjunto de D baseado no a_melhor;
    foreach D_v do
        Árvore_v = J48(D_v);
        Fixe a Árvore_v no galho obtido no passo anterior;
    end
end
return [Árvore]
```

O algoritmo *J48* adota a estratégia pós-poda, ou seja, faz a poda somente quando toda a árvore de decisão já foi gerada. Podar uma árvore, neste contexto, significa reduzir algumas sub-árvores a folhas, ou de outra forma, um ramo da árvore, a partir de determinado nó é cortado (transformado em folha). O corte de um ramo da árvore é guiado por um teste estatístico que tem em conta os erros em um nó e a soma dos erros nos nós que descendem desse nó. Assim, para cada nó, a poda só se concretiza se o desempenho da árvore não diminuir significativamente. Para além do problema do superajuste (*overfitting*).

2.2 Estado-da-Arte

O trabalho de Wu et al. (2008) apresenta um "ranking" dos 10 algoritmos mais influentes em Mineração de Dados identificados e eleitos na IEEE International Conference on Data Mining (ICDM) em dezembro de 2006. Os algoritmos que fazem parte do ranking são: *C4.5*, *k-Means*, SVM, *Apriori*, *EM*, *PageRank*, *AdaBoost*, *kNN*, *Naive Bayes*, e *CART*, estes 10 algoritmos cobrem as tarefas de classificação, clusterização, aprendizagem de máquina, regras de associação, mineração de links, que estão entre os tópicos mais importantes na pesquisa e desenvolvimento de Mineração de Dados.

No trabalho os autores fornecem ainda uma descrição de cada algoritmo e discutem

seu impacto, além de fornecerem uma revisão das pesquisas existentes sobre o algoritmo. A importância do trabalho para a área está ligada principalmente pela grande qualidade da conferência ICDM e dos membros convidados para eleição dos algoritmos mais relevantes para a área de Mineração de Dados. A seguir uma breve descrição de cada algoritmo:

- *C4.5*: é um algoritmo descendente direto dos algoritmos *CLS* (*Concept Learning System*) (HUNT; MARIN; STONE, 1966) e *ID3* (*Iterative Dichotomiser 3*) (QUINLAN et al., 1979) que são classificadores baseados em árvores de decisão, mas que também podem construir classificadores baseados em conjuntos de regras. Os aspectos da geração de árvores de decisão são apresentados na subseção 2.1.5, vale ressaltar que o algoritmo *J48* é a implementação em Java para o Weka do algoritmo *C4.5*. Um dos grandes feitos do algoritmo foi a introdução de um formalismo alternativo para geração de conjuntos de regras que consiste em uma lista de regras na forma "se A e B e C e... então classe X", onde as regras de cada classe são agrupadas. Uma instância é então classificada por encontrar uma regra que satisfaça as condições, se nenhuma regra é satisfeita, a instância é atribuída a uma classe de padrão. A principal desvantagem do conjunto de regras do *C4.5* está no uso do processador e memória que eles exigem. Para árvores de decisão, passando de 10 para 100 mil instancias aumentou o tempo de processamento em um computador de 1.4s para 61s, um fator de 44. O tempo necessário para conjuntos de regras, no entanto, aumentou de 32s para 9.715s, um fator de 300.

- *k-means*: é um algoritmo iterativo simples que particiona um determinado conjunto de dados em um número k de agrupamentos (*clusters*) aleatóriamente. Após essa etapa é calculada a similaridade de cada instância restante com o centro de cada agrupamento. A instância é então inserida no agrupamento com a menor distância, ou seja, maior similaridade. O centro do agrupamento é recalculado a cada nova instância inserida. Uma das limitações do algoritmo *k-means* é que ele se aplica apenas a casos onde as "t-uplas" de dados são completamente numéricas, dessa forma, os dados nominais precisam ser pré-processados para que o *k-means* seja aplicado sobre o conjunto de dados em questão.

- *SVM*: as Máquinas de Vetores de Suporte (*SVM*) oferecem um dos métodos mais robustos e precisos entre todos os algoritmos conhecidos. Tem uma base teórica sólida, requer apenas uma dúzia de exemplos para treinamento e é insensível ao número de dimensões. Além disso, métodos eficientes para a formação de *SVM* também estão sendo desenvolvidos em um ritmo crescente. O algoritmo consiste em determinar o hiperplano que separa duas classes e um conjunto de pontos a que elas pertencem, de forma a colocar o maior quantidade possível de pontos da mesma classe lado a lado, ao mesmo tempo que a distância de cada classe nesse

hiperplano é maximizada. Normalmente, as máquinas de vetor suporte operam em um espaço de dimensão maior que a dimensão dos dados originais, no chamado espaço de características. E é neste espaço que se consegue a maximização desejada para obter uma melhor capacidade de generalização de classificação. Atualmente um dos problemas da técnica de *SVM* é o tempo utilizado no aprendizado de máquina. Muitas pesquisas tem se concentrado neste aspecto.

- *Apriori*: dentre as abordagens de Mineração de Dados uma das mais populares é encontrar conjuntos de itens frequentes de um conjunto de dados de transação e derivar regras de associação. Essa tarefa de encontrar conjuntos de itens frequentes não é trivial, por causa da sua explosão combinatória. Uma vez que conjuntos de itens frequentes são obtidos, é simples gerar regras de associação. O *Apriori* é um algoritmo seminal para localizar conjuntos de itens frequentes usando a geração de candidatos. É caracterizado como um algoritmo de pesquisa que utiliza o princípio da antimonotonicidade do suporte de conjuntos de itens, ou seja, "se um conjunto de itens não é frequente, qualquer conjunto do seu superconjunto inicial nunca é frequente". Um pseudoalgoritmo para o *Apriori* pode ser visto em algoritmo 1. Uma limitação do algoritmo *Apriori* é o não processamento de váriaveis númericas, problema que pode ser resolvido através da etapa de transformação na fase do pré-processamento.

- *EM*: o algoritmo *EM* (*Expectation Maximization*), ou Maximização de Expectativas é um método, baseado em modelos, para estimar funções de máxima verossimilhança a partir de dados incompletos. Tem como base a ideia de substituir uma difícil maximização da verossimilhança por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original. Para isso ocorre diversas iterações que envolve dois passos no qual o primeiro consiste em estimar os dados que faltam para completar a amostra de dados incompleta. Esta estimativa é feita usando os valores das variáveis que foram observadas. O passo seguinte trabalha com as novas frequências esperadas obtidas anteriormente, aplica-se um algoritmo com dados completos. Sendo assim esses novos parâmetros substituem os parâmetros anteriores. A partir desses novos parâmetros o primeiro e o segundo passo são executados novamente, sendo repetidos até a convergência. Portanto o ponto forte do algoritmo *EM* é trabalhar em casos onde se tem conjuntos de dados com informações incompletas, o resultado é o agrupamento das instâncias de modo não supervisionado. Uma das limitações é que nem sempre a convergência é rápida, porque ela não é linear; dependendo do modelo, são usados processos numéricos para acelerá-la, por exemplo, o método *Newton-Raphson* combinado com o *EM*.

- *PageRank*: o algoritmo *PageRank* foi apresentado e publicado por Sergey Brin e Larry Page na sétima conferência internacional de *World Wide Web* (WWW7) em abril de

1998. É um algoritmo de classificação de pesquisa usando *hiperlinks* na Web. Baseado no algoritmo, construíram o motor de busca Google, que tem sido um enorme sucesso. O *PageRank* produz um ranking estático de páginas da Web, calculando o valor para cada página "offline", sendo assim ele não depende de consultas de pesquisa (queries). O algoritmo baseia-se no caráter democrático da Web usando links como indicadores de qualidade de uma página. Em suma, o *PageRank* interpreta um hiperlink de uma página X para uma página Y como um voto. Entretanto o algoritmo, análisa mais do que apenas o número de absoluto de votos, ou de links que uma página recebe, ele também análisa a página que realiza o voto. Votos realizados por páginas que os autores consideram importantes recebem um peso maior e ajudam a tornar outras páginas mais "importantes". Esta é exatamente a ideia de classificação por prestígio em redes sociais. O algoritmo tem mudado bastante desde a sua concepção e hoje pesquisadores propuseram modelos alternativos e melhorias para sua computação, por exemplo, adicionando a dimensão temporal, etc.

- *AdaBoost*: é um algoritmo do tipo *boosting* que funciona em conjunto com outros algoritmos de aprendizado de máquina na tentativa de prever um comportamento baseando-se na observação de dados, tornando-os mais sensíveis aos valores estatísticos discrepantes. O *AdaBoost* treina classificadores, sendo que os padrões usados nos treinamentos têm pesos associados, que dão importância maior ou menor para o padrão no treinamento. Inicialmente todos os padrões têm peso igual, mas quando um é classificado incorretamente, ele tem seu peso aumentado, se tornando mais relevante para a próxima iteração. Pelo *AdaBoost*, os falsos positivos de um classificador, por terem peso maior, são considerados mais relevantes para o próximo, diminuindo a quantidade total de falsos positivos. Por estarem em cadeia, cada um classificando corretamente todos os padrões da classe desejada, com taxa de falsos positivos menor que 50% e dando ênfase aos falsos positivos do classificador anterior, a taxa de verdadeiros positivos é de 100% e a de falsos positivos tende a ser muito baixa. As vantagens do *AdaBoost* estão na simplicidade de implementação, flexibilidade para trabalhar em vários tipos de problemas de classificação e baixo valor computacional já que funciona de forma linear. Uma das limitações é em relação a ruídos nos dados, devido ao princípio do algoritmo de enfatizar os dados mais difíceis de serem classificados, os ruídos podem ser enquadrados nesse grupo e assim prejudicar os resultados gerados. O pré-processamento então tem um papel decisivo no aumento da qualidade dos dados, eliminando ruídos e aumentando a qualidade dos resultados.

- *kNN*: também conhecido por k-vizinhos mais próximos (*k-nearest neighbor*), esse algoritmo reconhece padrões na localização de um dado e associa esse dado a um identificador. O algoritmo procura um grupo k de objetos no conjunto de

treinamento que estão mais próximos (por distância ou métrica de similaridade) ao novo objeto submetido a classificação. A nova instância é classificada na classe mais comum entre todos os k registros mais próximos. Portanto são três elementos chaves utilizados pelo *kNN* um conjunto de objetos já rotulados, uma distância ou métrica de similaridade entre objetos e o valor de k, correspondente ao número de vizinhos próximos considerados para o cálculo da distância e classificação. Uma das limitações dessa abordagem é a escolha do valor de k que quando é pequeno demais, o resultado do algoritmo pode ficar sensível a valores discrepantes (*outliers*). Por outro lado, se o valor de k for muito grande então a vizinhança pode incluir objetos pertencentes a classes distintas. Nesse caso o pré-processamento pode ser utilizado para amenizar ou até mesmo remover *outliers* e ajudar na escolha de um valor ideal para k.

- *Naive Bayes*: é um método que usa probabilidade condicional baseada no teorema de Thomas Bayes. Segundo o teorema de Bayes, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu: $Probabilidade(B\ dado\ A) = Probabilidade(A\ e\ B)/Probabilidade(A)$. Este método é importante por várias razões já que é muito fácil de construir, não necessita de qualquer parâmetros iterativos complicados e esquemas de estimativa. Isto significa que pode ser facilmente aplicado a grandes conjuntos de dados e é fácil de interpretar. Ele pode não ser o melhor classificador possível em um determinado caso, mas ele geralmente executa bem a tarefa de classificação e é robusto o bastante. Um grande número de modificações foram introduzidas por áreas como a Estatística, Mineração de Dados e por comunidades de reconhecimento de padrões, na tentativa de torná-lo mais flexível, mas é preciso reconhecer que tais modificações são complicações que desvirtuam sua simplicidade básica. O algoritmo *Naive Bayes* não trata dados númericos contínuos, ou seja, para aplicação do algoritmo, os dados precisam passar pela etapa de transformação do pré-processamento onde serão discretizados.

- *CART*: o nome do algoritmo significa Classificação e Árvore Regressiva (*Classification and Regressive Tree*). A árvore de decisão *CART* é um procedimento binário e recursivo capaz de processar atributos nominais e númericos. O algoritmo realiza uma divisão binária resultando em dois nós filhos e incorpora uma medida de impureza chamada Gini (BREIMAN et al., 1984) para estas divisões, onde Gini é uma medida de dispersão que depende da distribuição das categorias resultantes. O valor do Gini varia de 0 a 1 assumindo o valor máximo (pior caso) no caso de distribuições equilibradas e o resultado mínimo (melhor caso) quando todos os registros de um nó estão concentrados em uma única categoria. Têm-se dois fatores principais no algoritmo de partição, a forma para selecionar uma divisão para cada nó intermediário (crescimento) e uma regra para determinar quando um nó é terminal. O

problema chave no algoritmo de partição recursiva é a confiabilidade das estimativas do erro usado para selecionar as divisões. As escolhas da divisão em níveis maiores da árvore produzem, frequentemente, estatísticas não-confiáveis apesar da estimativa do "erro de resubstituição" (estimativa obtida com os dados de treinamento usado durante o crescimento da árvore) manter-se decrescendo. Com isto, a precisão das estimativas do erro é fortemente dependente da qualidade da amostra, sendo assim o pré-processamento representa um papel fundamental no aumento da qualidade da amostra através da realização das etapas de limpeza, transformação, redução e integração dos dados.

Ressalta-se que a conferência ICDM tem publicado os melhores trabalhos submetidos a conferência ao longo dos anos, e vários dos algoritmos acima citados para classificação, clusterização, aprendizagem estatística e análise de associação foram selecionados pelos comitês do programa da ICDM dos anos anteriores, para publicação na conferência após suas revisões e expansões. Assim sendo, eles são considerados o estado-da-arte em Mineração de Dados.

2.3 Trabalhos Importantes para Área

Neste item é feita uma revisão bibliográfica, com um breve relato de outros trabalhos desenvolvidos com a utilização de diferentes técnicas em Mineração de Dados, em especial trabalhos que apresentem o pré-processamento de dados como um passo crítico em um projeto de Mineração de Dados.

No trabalho de Fayyad et al. (1996), os autores descrevem as relações entre KDD e Mineração de Dados e como estão relacionados à estatísticas e ao aprendizado de máquina. Eles conceituam o KDD como todo o processo de descoberta de conhecimento e a Mineração de Dados como uma etapa neste processo, assim como o pré-processamento. No trabalho são relatadas técnicas específicas para Mineração de Dados tais como árvore de decisão, regressão não-linear e modelos de aprendizagem relacional. É discutido que não existe um método mais eficiente que sirva para todas as aplicações. A escolha do método depende do objetivo da Mineração de Dados e dos dados em si.

A Mineração de Dados tem aplicações muito famosas em pesquisas cientificas como o software SKICAT (*Sky Image Cataloging and Analysis Tool*) na área de astronomia. O SKICAT ajuda astrônomos a descobrir quasares distantes. É a primeira ferramenta de destaque na Mineração de Dados, especialmente em aplicações em astronomia e ciência espacial. Usando SKICAT, os astrônomos descobriram 16 novos quasares extremamente distantes (WEIR et al., 1995).

A aplicação de Mineração de Dados em biologia molecular está concentrada prin-

cipalmente em engenharia genética. Como já foi sugerido por Zhu e Davidson (2007), existem 10.000 proteínas diferentes, o método de análise de séries de moléculas biológicas e a tecnologia de pesquisa em banco de dados de genes têm contribuído grandemente para muitas descobertas importantes e mostrado como técnicas de pré-processamento podem reduzir o número de características a serem analisadas pelos algoritmos levando a uma maior precisão da predição de novas características.

Trabalhos sobre o uso da Web como o de Lakshmi, Rao e Reddy (2013), utilizam a Mineração de Dados em um caso especial chamado Mineração dos Registros de Navegação na Web (*Web Usage Mining*) para extração de conhecimento útil a partir de *logs* de servidores web, registros de bancos de dados, consultas de usuários, *cookies* e perfis a fim de analisar o comportamento do usuário. É demonstrado ao longo do trabalho que a análise desses dados requer abstração para a descoberta de padrões e que essa abstração pode ser conseguida através do pré-processamento de dados, essa demonstração é feita utilizando alguns *logs* de servidores e utilizando técnicas de pré-processamento específicas para cada caso como a limpeza e a transformação dos dados.

Como último exemplo apresentado de trabalho importante para a área, há o artigo de Márquez-Vera et al. (2013), onde eles publicaram um trabalho que tinha como objetivo prever a reprovação escolar usando um conjunto de dados escolares. A estratégia utilizada foi acrescentar uma etapa de pré-processamento para limpar e integrar os dados, obtendo assim os melhores atributos para a resolução do problema da alta dimensionalidade. No comparativo realizado, os autores demonstraram o ganho de precisão das regras de classificação alcançado com o uso do pré-processamento.

3 Metodologia

Este capítulo apresenta o método de pesquisa, equipamentos, softwares, sistema operacional, além dos processos aplicados sobre os conjuntos de dados.

3.1 IRPM

O presente trabalho utilizou como metodologia de pesquisa o IRPM (*Interdisciplinary Research Project Management*) (CAETANO, 2012) e (LETOUZE et al., 2012), o texto a seguir foi extraído na íntegra com autorização do autor.

De acordo com Letouze (2011), o IRPM (*Interdisciplinary Research Project Management*, Gerência de Projetos de Pesquisa Interdisciplinar) é uma estratégia de abordagem interdisciplinar para um problema real, usando de Gerenciamento de Projetos conceitos propostos por Pierrakos, Zilberberg e Anderson (2010) e um problema de abordagem baseada na aprendizagem, que é mostrado por Savery (2006) em seu artigo relacionado. Através da Figura 4 se tem noção das etapas do IRPM que serão explicadas a seguir.

Em IRPM, a fase de **iniciação** começa com a escolha do problema real para resolver e identificar pelo menos dois campos para uma abordagem interdisciplinar. Estes campos são necessários para: documentar as restrições problemas reais e suposições; definir a estratégia, identificar os critérios de desempenho, determinar as necessidades de recursos; definir orçamento; e produzir documentação formal.

O **planejamento** por sua vez é o refinamento do projeto que analisa o problema real através do uso dos campos escolhidos. Estes estudos podem produzir uma nova metodologia fundamental. Portanto, em fase de execução, mesmo se novos conceitos não são obtidos, um material didático pode ser preparado e utilizados em sala de aula para uma abordagem Baseada em Problemas de Aprendizagem.

Figura 4 – A Investigação Interdisciplinar modelo de Gestão de Projeto. (Fonte: Traduzida de (LETOUZE, 2011))

Explicando de uma forma detalhada cada etapa do IRPM temos:

1. Iniciação: para determinar os objetivos do projeto, as entregas e saídas de processos; para documentar projeto restrições e suposições; para definir a estratégia; identificar os critérios de desempenho; para determinar necessidades de recursos; para definir o orçamento e para produzir uma documentação formal.

2. Planejamento: para refinar projeto; para criar uma estrutura de divisão de trabalho; para desenvolver o recurso plano de manejo; para aperfeiçoar estimativas de tempo e de custos, estabelecer controles de projeto; para desenvolver o plano de projeto e obter a aprovação do plano.

3. Execução: a comprometer recursos; para implementar recursos; para gerenciar o progresso; para se comunicar progredir e aplicar procedimentos de garantia de qualidade.

4. Controle: para medir o desempenho; para refinar os limites de controle; para tomar ações corretivas; para avaliar eficácia das ações corretivas; para garantir a conformidade plano; a reavaliar os planos de controle; para responder a evento de risco e para monitorar a atividade do projeto.

5. Encerramento: para obter a aceitação das entregas; documentar as lições aprendidas; para facilitar o fechamento; preservar os registros de produtos, ferramentas para liberar recursos.

3.1.1 Aplicação do IRPM

De acordo com a Figura 4, como princípio deste trabalho, na etapa de iniciação escolheu-se o problema de encontrar relacionamento entre fatores abióticos e bióticos do ictioplâncton (conjunto dos ovos e larvas de peixes que apresentam um comportamento planctônico) de um reservatório de água doce da Amazônia Legal como problema real a ser tratado. Deste modo foram identificados as seguintes áreas a serem abordadas: biodiversidade, com estudo do ictioplâncton da região; Mineração de Dados e seus algoritmos; pré-processamento devido os dados estarem em planilhas diferentes, dados ausentes, etc.

Em seguida, na fase de planejamento, estudou-se o problema onde se constatou que o impacto ambiental da exploração humana motivou esforços no sentido do monitoramento do ictioplâncton nas fases iniciais de desenvolvimento gerando uma das primeiras bases sistematizadas de informação do grupo de pesquisa da Amazônia Legal, com dados coletados antes, durante e após o enchimento do reservatório. No caso do pré-processamento, percebeu-se a necessidade de melhorar a qualidade dos dados ao experimentar a aplicação dos algoritmos de Mineração de Dados sobre o conjunto de dados formado pelo grupo de pesquisa, onde constatou-se que isto não seria possível sem um processamento dos dados em uma etapa anterior a fase da Mineração de Dados.

Portanto, a metodologia para estudar o problema real foi proposta no artigo "*Exploring an Ichthyoplankton Database from a Freshwater Reservoir in Legal Amazon*" (Aplicação do algoritmo Apriori para uma base de dados de ictioplâncton em um reservatório de água doce da Amazônia Legal) (SILVA et al., 2013) e é apresentado, em versão traduzida para o português, no Capítulo 4. Observa-se que não foram aplicadas todas as etapas possíveis do IRPM. Por exemplo, na fase de execução poderia-se elaborar um material didático utilizando PBL - *Problem Based Learning* (Aprendizado Baseado em Problemas), ou desenvolver uma nova tecnologia que seria implantada "in loco" para a monitoração dos dados abióticos de modo a alimentar uma base de dados projetada em concordância com os resultados obtidos no artigo (SILVA et al., 2013). Em seguida ainda na fase de execução, a aplicação desta nova tecnologia daria suporte as pesquisas do grupo de pesquisa da Amazônia Legal. Já na fase de controle, teríamos o levantamento dos parâmetros para estudo e pesquisas adicionais que provavelmente culminariam em publicações científicas - fase de encerramento.

Em uma segunda pesquisa ainda de acordo com a metodologia proposta por Letouze (2011), escolheu-se o problema de identificação dos indicadores sociais que levam ao trabalho infantil e, dessa maneira, foram identificadas as seguintes áreas a serem abordadas: área social, com dados de programas de programa de assistência social do Governo Federal; Mineração de Dados e seus algoritmos; pré-processamento devido os dados estarem tabelas de banco de dados diferentes, dados ausentes, etc.

Na sequência, um estudo do problema foi realizado sobre as características de inclusão de indivíduos e famílias no Programa de Assistência Social Cadastro Único (CadÚnico) do Governo Federal e sobre o conjunto de dados oriundos do cadastro de indivíduos que recebem a assistência do governo. Aproveitando a metodologia desenvolvida no trabalho anterior (SILVA et al., 2013) um novo estudo sobre os algoritmos de Mineração de Dados foi realizado na busca por métodos que proporcionariam resultados com maiores índices de precisão perpassando pelas etapas de pré-processamento, novamente na tentativa de melhoria da qualidade dos dados.

A metodologia para estudar o problema foi proposta no artigo "*Exploring Social Data to Understand Child Labor*" (Explorando dados sociais para entender o trabalho infantil) (RODRIGUES; PRATA; SILVA, 2013) e é apresentado, em versão traduzida para o português, no Capítulo 5.

O terceiro projeto de pesquisa onde o IRPM foi aplicado, durante a fase de iniciação escolheu-se o problema de utilizar a Mineração de Dados como ferramenta para a realização da Análise Exploratória em pesquisas ecólogicas. Dessa maneira, foram identificadas as seguintes áreas: Ecossistemas, com exemplos de pesquisas desenvolvidas utilizando estátistica tradicional na modelagem dos problemas; Mineração de Dados para a modelagem algorítmica; pré-processamento como etapa fundamental na melhoria da qualidade dos resultados; a Análise Exploratória como um tipo de análise a ser empreendida antes mesmo de que as hipóteses sejam formuladas, ou quando o conhecimento sobre o domínio é escasso.

Durante a fase de planejamento o problema foi estudado permitindo o desenvolvimento de uma abordagem alternativa às técnicas de estatística tradicionais na modelagem do problema. Durante essa fase experiências adquiridas nas pesquisas anteriores tiverem papel fundamental no desenvolvimento do projeto. A área de Mineração de Dados foi estudada a fim de extrair técnicas que possibilitassem o suporte a Análise Exploratória; o pré-processamento e as etapas que ajudam a aumentar a qualidade, dos dados e consequentemente, dos resultados da Mineração de Dados.

A metodologia proposta para realização da análise exploratória através do uso de Mineração de Dados foi proposta no artigo Mineração de Dados como ferramenta de análise exploratória em Sistemas Ecológicos e é apresentado no Capítulo 6. Nota-se que, neste caso também, nem todas as fases do IRPM foram adotadas.

3.2 Materiais

Nesta seção, apresenta-se o material utilizado no desenvolvimento da pesquisa, tais como equipamentos, softwares e sistema operacional.

3.2.1 Equipamentos

Para os experimentos foi utilizado o computador MacBook Pro Core i7 (I7-2675QM) a 2,2 GHz 64bits com 4 GB de memória DDR3 a 1333 MHz, placa de vídeo AMD Radeon HD 6750M com comutação automática de gráficos, com memória gráfica de 512MB GDDR5 e disco rígido de 500GB Serial ATA, 5400 rpm.

3.2.2 Softwares

Neste trabalho foi utilizada uma implementação do algoritmo *Apriori* para a base de dados de ictioplâncton e *J48* para a base de dados sociais do Governo Federal com o uso da ferramenta *Waikato Environment for Knowledge Analysis* (Weka) na versão 3.6.9 (HALL et al., 2009).

A ferramenta Weka é de código aberto sob licença GNU *General Public License* foi desenvolvida pela Universidade de Waikato na Nova Zelândia em linguagem de programação Java (ORACLE, 2013) e implementa uma grande variedade de métodos de Mineração de Dados. Ela disponibiliza também diversos algoritmos de pré-processamento de dados, bem como de análise resultados que podem ser acessados através da interface gráfica ou por linha de comando. A ferramenta possui seu próprio formato para arquivos utilizados como entrada, o *Attribute Relation File Format* (ARFF), mas outros formatos também podem ser utilizados, como por exemplo CSV (Valores Separados por Vírgula), além de ser possível conectar a um banco de dados *Structured Query Language*, ou Linguagem de Consulta Estruturada (SQL) diretamente utilizando o JDBC (*Java Database Connectivity*). Ainda existe a opção de utilizar o Weka através de sua API *Application Programming Interface*, ou Interface de Programação de Aplicativos incluindo o JAR (*Java ARchive*) do Weka no projeto desejado.

Apesar do Weka fornecer algumas opções de visualização de resultados optou-se pela utilização do programa R versão 3.0.1 para geração das visualizações por apresentar mais recursos nessa etapa. O R é uma linguagem e ao mesmo tempo um ambiente para computação estatística e construção de gráficos. Como um projeto GNU ele é um *software* livre que pode ser baixado e distribuído gratuitamente de acordo com a licença GNU (TEAM, 2013). Ele foi desenvolvido como uma derivação da linguagem estatística S (BECKER, 1994), inicialmente para ambientes Unix/Linux, embora atualmente esteja disponibilizado em diversas plataformas, tais como o Windows ou MacOS. Ele fornece uma ampla variedade de estatísticas como modelagem linear e não linear, testes estatísticos clássicos, análises de séries temporais, classificação, agrupamento, etc., bem como técnicas gráficas para apresentação de resultados. Tem por paradigma a programação orientada a objetos, permitindo que novas características sejam acrescentadas através de pacotes (*packages*). O programa proporciona a importação de dados reconhecendo vários formatos, incluindo o CSV. O programa R como ambiente ainda inclui:

- um tratamento eficaz dos dados e facilidade de armazenamento;
- um conjunto de operadores para cálculos em vetores e matrizes;
- uma coleção grande, coerente e integrada de ferramentas intermediárias para análise de dados;
- recursos gráficos para análise de dados e uma linguagem de programação bem desenvolvida, simples e eficaz, que inclui condicionais, laços, funções recursivas definidas pelo usuário e recursos de entrada e saída.

A versão do Java JDK (Java Development Kit) utilizada nos experimentos foi 7 update 25 (7u25).

3.2.3 Sistema Operacional

Como sistema operacional do computador foi utilizado o OS X Mountain Lion versão 10.8.3.

3.3 Processos

O processo adotado no presente trabalho segue o fluxograma ilustrado na Figura 5. Com a definição do problema a ser atacado, o primeiro passo do processo diz respeito ao a fase de pré-processamento que inicia com as seguintes etapas em sequência:

1. Integração dos conjuntos de dados com a finalidade de construir um conjunto de dados único e consistente. As chaves amostrais ou primárias são utilizadas para resolver conflitos e divergências.
2. Limpeza dos dados, registros incompletos são eliminados, campos com valores errados e dados inconsistentes são corrigidos ou eliminados conforme a técnica utilizada para tal, como por exemplo a substituição de dados ausentes pelo sinal de "?" que para ferramenta Weka sinaliza um campo faltante e será analisado como tal. Valores estatísticamente discrepantes são encontrados utilizando agrupamento dos registros e assim descartados.
3. Redução das variáveis presentes no conjunto de dados para evitar problemas com a alta dimensionalidade, utilizando algoritmos específicos para esse fim como o *Cfs-SubsetEval* apresentado anteriormente na subseção 2.1.3, permitindo maior eficiência dos algoritmos na fase de Mineração de Dados.

4. Transformação de dados seja por estarem errados (como por exemplo formatos de datas) ou como pré-requisito para a execução de algum algoritmo de Mineração de Dados que não suporte um tipo ou outro de atributo. A discretização dos dados pode ser utilizada para esse fim. Outras técnicas utilizadas são a normalização e o agrupamento de atributos.

Uma vez finalizada a fase de pré-processamento um novo conjunto de dados é formado e é sobre este conjunto final que é realizada a modelagem algorítmica, com a aplicação de algoritmos de Mineração de Dados sobre o conjunto de dados. Nesta fase os algoritmos são escolhidos de acordo com o problema e os dados do conjunto final, entre o universo de algoritmos disponíveis se exclui aqueles que não fornecem os resultados esperados ou que não se aplicam e é dada a preferência por algoritmos de Mineração de Dados recomendados para aquele tipo de problema, seja ele de classificação, agrupamento (*clustering*), regras de associação, etc., por especialistas da própria área de Mineração de Dados. Uma vez escolhido o algoritmo e definido os valores das métricas de avaliação, o algoritmo é aplicado naquele conjunto de dados e o modelo é gerado. Se algum algoritmo indicado pelo especialista não puder ser utilizado por restrição ao tipo do dado (nominal ou númerico), ou ainda, se a precisão ou outra métrica especificada pelo pesquisador estiver abaixo do estabelecido no passo de parametrização das métricas de avaliação, pode-se retornar para a fase de pré-processamento para ajustes no conjunto de dados para que o mesmo seja novamente submetido a modelagem algorítmica.

Com o modelo gerado pela modelagem algorítmica inicia-se a fase do pós-processamento que será responsável por facilitar o processo de interpretação dos resultados. Sobre o modelo é gerada a visualização dos resultados de acordo com o tipo de algoritmo utilizado, nesse caso essas visualizaçãos podem ser do tipo árvores de decisão, conjuntos de regras, gráficos de dispersão, etc..

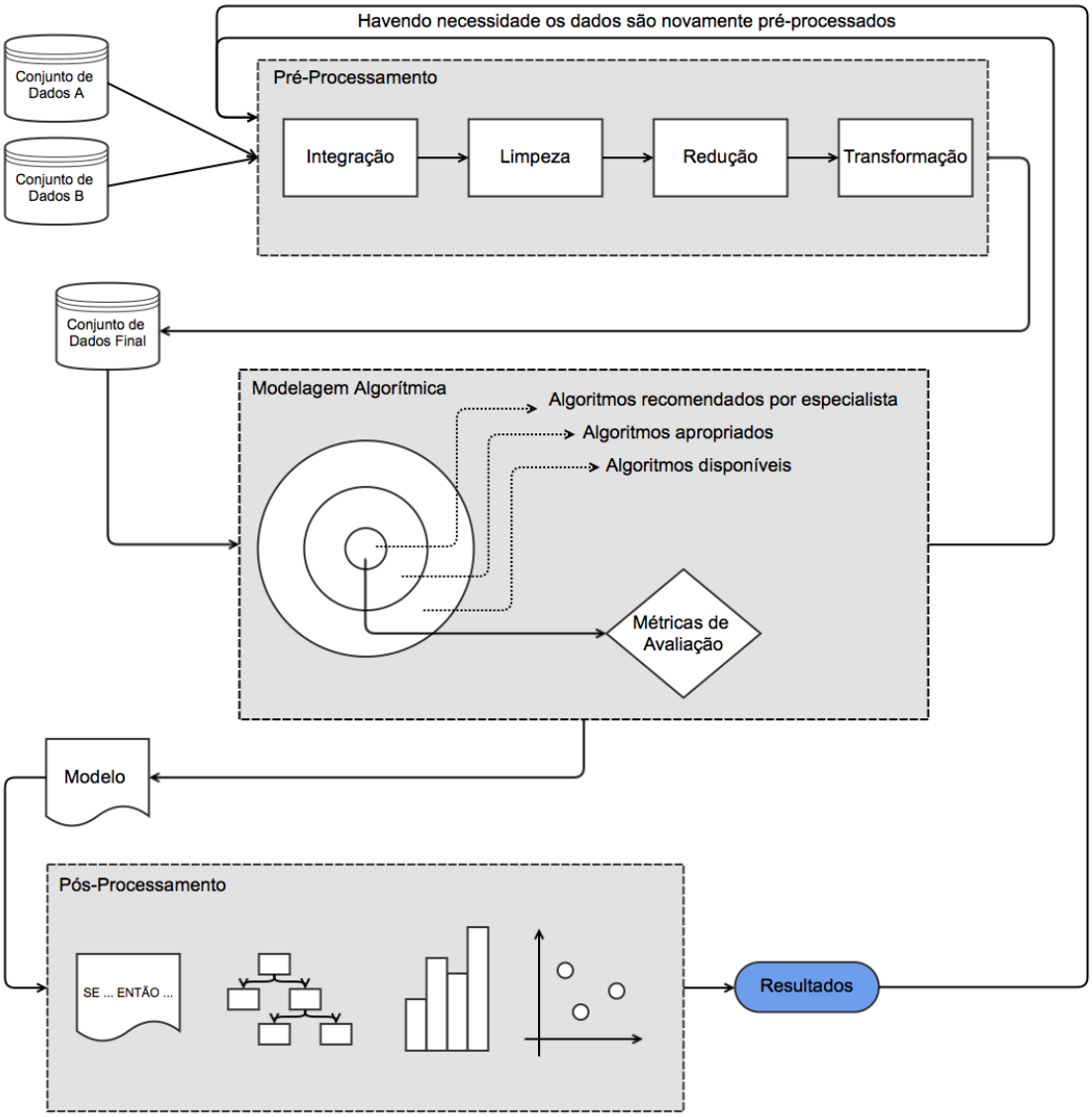

Figura 5 – Diagrama do processo utilizado no presente trabalho.

Com a visualização dos resultados pode-se a qualquer momento retornar para a fase de pré-processamento na tentativa de melhora-los. Dessa forma, ressalta-se a interatividade do processo com a participação do pesquisador durante as fases avaliando os resultados e decidindo quando encerrar ou não o processo.

4 Aplicação do algoritmo Apriori para uma base de dados de ictioplâncton em um reservatório de água doce da Amazônia Legal

Michel de A.Silva Daniela Queiroz Trevisan David N. Prata
Elineide E. Marques Marcelo Lisboa Monica Prata

Abstract

The purpose of this study is the use of data mining techniques for exploratory analysis in a database of ichthyoplankton from a reservoir of freshwater at Legal Amazon. This same database had already been analyzed using statistical techniques. However, the analysis was unsuccessful in finding a relationship between biotic and abiotic factors. Application of Apriori algorithm allowed the generation of association rules that yielded the discovery of unique knowledge for understanding the process of fish spawning in Tocantins River. In this case, we have demonstrated the effective use of data mining for the discovery of patterns and processes in ecological systems, suggesting that statistical methods often used by ecologists can be coupled with data mining techniques to generate hypotheses.

Key-words: data mining. Apriori. exploratory analysis. ichthyoplankton.

Resumo

O propósito deste estudo é apresentar o uso de técnicas de mineração de dados para análise exploratória em uma base de dados de ictioplâncton de um reservatório de água doce da Amazônia Legal. Esta mesma base de dados já havia sido analisada utilizando técnicas estatísticas, porém, sem sucesso na busca de uma relação entre dados bióticos e abióticos. A aplicação do algoritmo Apriori permitiu a geração de regras de associação que proporcionaram a descoberta de conhecimento singular para o entendimento do processo de desova dos peixes na bacia do rio Tocantins. Neste caso, conseguimos demonstrar o uso eficaz da mineração de dados para o descobrimento de padrões e processos em sistemas ecológicos, sugerindo que os métodos estatísticos frequentemente usados pelos ecologistas podem ser aliados às técnicas de mineração de dados para a geração de hipóteses.

Palavras-chaves: mineração de dados. Apriori. análise exploratória. ictioplâncton.

1 Introdução

Nas últimas décadas, com a evolução das tecnologias de hardware, a quantidade de dados armazenados aumentou sistematicamente. Esses dados têm sido mantidos em fontes variadas, tais como sistemas de bancos de dados, páginas Web, arquivos convencionais, planilhas eletrônicas, *tablets* e *smartphones*.

A tecnologia da informação, especificamente a mineração de dados, tem permitido extrair informação e maximizar os resultados obtidos com aplicação de diversas técnicas de inteligência artificial. A manipulação dessas bases de dados culminou em novas formas de relacionar essa informação que expandem a discussão do tema e suscitam a elaboração de novas questões de estudo.

Com o aumento da população mundial e os avanços tecnológicos nos processos produtivos, vivemos em um mundo "dirigido à recursos". O impacto ambiental da exploração humana é um tema atual que tem rogado esforços por parte da comunidade científica. A mineração de dados ecológicos busca resposta às questões de como podemos entender e utilizar, de forma mais eficiente nossos recursos naturais.

Ecologistas têm utilizado principalmente métodos estatísticos para análise das relações entre uma resposta observada e um conjunto de preditivas em um conjunto de dados. Esta abordagem para análise de dados é mais apropriada para teste de hipóteses. Quando o conhecimento anterior for mínimo e as hipóteses não estão claramente desenvolvidas, as análises exploratórias são mais apropriadas do que análises confirmatórias (HOCHACHKA et al., 2007). As técnicas de mineração de dados são frequentemente mais poderosas, flexíveis e eficientes para análise exploratória do que as técnicas estatísticas ex.: (BREIMAN, 1996), (BAUER; KOHAVI, 1999), (HASTIE et al., 2009).

Neste trabalho apresentamos o processo de aplicação do algoritmo Apriori (AGRAWAL SRIKANT et al., 1994) de regras de associação, em busca por padrões sobre uma base de dados ecológica de ictioplâncton. A concepção deste processo possibilitou extrair conhecimento do domínio da base de dados que não puderam ser descobertos somente com a utilização de métodos estatísticos (AGOSTINHO et al., 2007). A aplicação desta técnica permitiu descobrir regras entre propriedades, gerando informações que não foram previstas, estando ocultas em meio aos dados.

O objetivo principal deste trabalho foi responder a seguinte pergunta: "*existe alguma relação entre os fatores abióticos com o estágio das larvas?*".

Na primeira fase, a modelagem do problema, definimos as questões para as quais se buscam respostas e as informações de significância de utilização de cada uma das propriedades. Na segunda fase, pré-processamento, selecionamos os dados a serem tratados, integrando, reduzindo e transformando-os com o objetivo de adicionar qualidade, tornando-os exequíveis. A partir deste ponto, terceira fase, iniciamos a fase de experimentos da mineração de dados. Onde, o algoritmo Apriori, foi aplicado sobre os dados pré-processados, com tratamento dos parâmetros de graus de confiança e suporte. A quarta fase, considerada pós-processamento, foi compreendida pela tarefa de refinamento e interpretação das regras extraídas, incluindo a avaliação de especialistas do domínio para análise dos resultados.

2 Base de Dados do IctioPlâncton

No contexto da conservação e gestão de recursos naturais, as bases de dados tendem a se tornar mais extensas com a ampliação das áreas de coleta, das séries temporais e do detalhamento das redes de amostragem. Atualmente, o monitoramento das modificações ocorridas nos sistemas naturais, antrópicas ou não, é um dos principais desafios à gestão ambiental em áreas tropicais. A detecção e descrição das alterações, a identificação de variáveis indicadoras de qualidade ambiental, e a construção e aferição de modelos para previsões futuras em sistemas complexos, como os ecossistemas aquáticos, apesar de desejáveis, estão em fase inicial de elaboração em áreas tropicais. A alta diversidade de espécies e de relações entre estas e os fatores ambientais possibilitam uma grande variação de cenários e interações que exigem técnicas variadas para serem identificadas.

Neste cenário, as informações obtidas durante o monitoramento da fauna de peixes, tanto nas fases iniciais de desenvolvimento (ovos e larvas) como na fase adulta, associados aos empreendimentos hidrelétricos constituem um conjunto de dados ainda não explorado.

Na bacia do rio Tocantins, assim como em outros rios do Brasil e do mundo, a fauna de peixes tem sido modificada sistematicamente pela construção de usinas hidrelétricas. Os sete empreendimentos em funcionamento no rio Tocantins tem modificado de forma acentuada o ambiente e a fauna de peixes, que pode estar ameaçada pela construção das demais usinas previstas para esta bacia (MME, 2012).

Contudo, acompanhar o comportamento da fauna de peixes e identificar os principais locais de desova e de desenvolvimento das fases iniciais, visando à delimitação de áreas de conservação, é uma das estratégias de conservação. Estas estratégias podem ser aperfeiçoadas e refinadas com o incremento do banco de dados, mas, necessariamente, requer a utilização das tecnologias de informação.

A utilização de diferentes técnicas de análise, mesmo quando a captura de dados está em andamento, auxilia também na criticidade da rede amostral, apontando lacunas que podem ser preenchidas com acertos na coleta das informações.

O monitoramento da fauna de peixes na área de influência da Usina Hidrelétrica do Lajeado, na bacia do Rio Tocantins, gerou uma das primeiras bases sistematizadas de informação do grupo de pesquisa na Amazônia Legal, com dados coletados antes, durante e após o enchimento do reservatório. Mudanças na distribuição e abundância dos peixes na fase larval têm sido relatadas utilizando metodologias tradicionais de análise (Agostinho et al., 2001). A utilização da mineração de dados pode agregar valor às pesquisas realizadas e indicar novas questões, metodologias e resultados.

3 Coleta de Dados

Os estudos foram conduzidos no reservatório da Usina de Lajeado (Usina Hidrelétrica Luís Eduardo Magalhães), Figura 1, cujo enchimento foi concluído em 2002.

A coleta de dados de larvas de peixes foi realizada mensalmente no período de abril de 2010 a março de 2012, em doze pontos localizados no reservatório (jusante, reservatório e montante) e na cabeceira e foz de seus principais afluentes (rios Santa Tereza, Santo Antônio; São Valério, Formiga, Crixás, Areias, Matança, Água Suja e Santa Luzia), utilizando a metodologia indicada em (NAKATANI, 2001). Ao mesmo tempo, foram registradas as informações referentes à temperatura do ar e da água, condutividade, pH e transparência

da água.

Os dados de temperatura da água (°C) foram obtidos utilizando-se um termômetro digital; a transparência da coluna da água (m), com disco de Secchi de 0,30m de diâmetro; o potencial hidrogeniônico (pH) e a condutividade elétrica ($\mu S.cm - 1$) por meio de potenciômetros digitais portáteis e as concentrações de oxigênio dissolvido através de oxímetros digitais portáteis (YSI, $mg.L - 1$).

As amostras foram tomadas por meio de arrastos horizontais, com rede de plâncton cônicocilíndrica, malha de 500 μm, comprimento de 1,5 m (parte cônica com 0,9 m e a cilíndrica 0,6 m), e um fluxômetro mecânico acoplado à boca da rede para obtenção do volume de água filtrada. As amostras foram tomadas à cerca de 20 cm de profundidade, no meio e nas proximidades das margens direita e esquerda dos locais com mais de 20 m de largura.

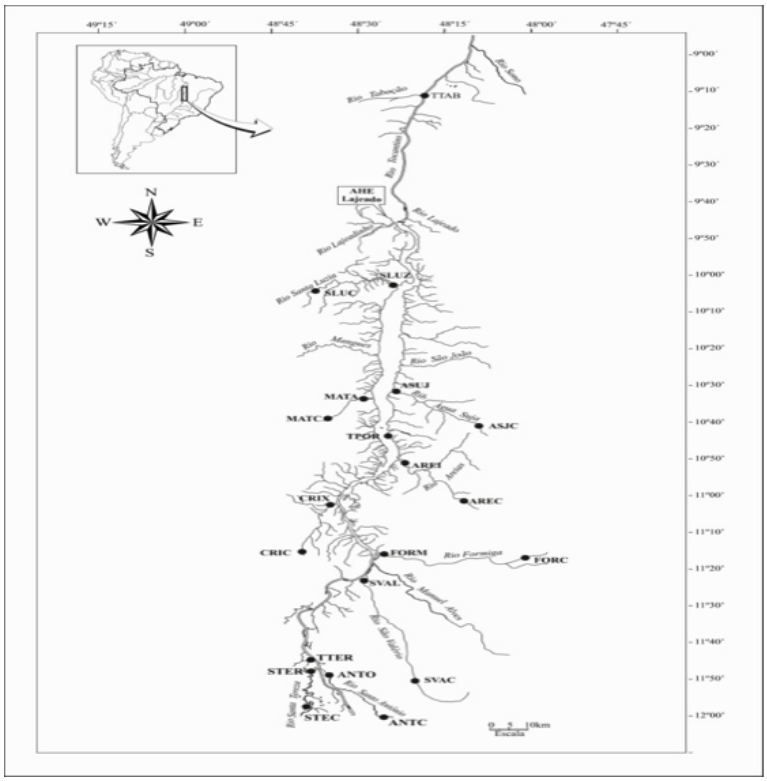

Figura 1 – Localização dos pontos de amostragem no reservatório do Lajeado e sua área de influência.

As amostras obtidas foram fixadas em formalina e analisadas em microscópio estereoscópico. As larvas foram separadas de acordo com o estágio de desenvolvimento em: larval vitelínica, pré-flexão, flexão e pós-flexão, considerando critérios como a presença do saco vitelínico, a abertura da boca e o desenvolvimento da nadadeira cauda e seus elementos de suporte, seguindo Nakatani (2001).

A identificação das larvas ocorreu em nível de ordem e família quando possível de acordo com Nakatani (2001), que considera os caracteres morfológicos como tamanho e posição dos olhos, forma e pigmentação do corpo, posição da abertura anal em relação ao corpo, sequência de formação e posição das nadadeiras, presença de barbilhões e de dados merísticos, como número de miômeros e de raios das nadadeiras. O material obtido está

depositado como material testemunho na Coleção de Peixes da Universidade Federal do Tocantins – UFT.

A abundância de ovos e larvas foi calculada em número por 10 metros cúbicos de água filtrada, conforme a metodologia de (Tanaka, 1973) modificada por Nakatani (2001).

Os dados obtidos foram organizados considerando as datas e os locais de amostragem, o ponto de realização da coleta (margem/meio; superfície), e armazenados em dois arquivos do tipo xls. Um dos arquivos com dados abióticos, que traz informações do ambiente da coleta, e outro arquivo com dados bióticos, que tem informações sobre os ovos e larvas coletadas.

4 Metodologia da Mineração dos Dados

A primeira etapa do trabalho consistiu no levantamento dos algoritmos de mineração de dados que encontrassem padrões entre as propriedades dos dados, identificando assim relacionamentos desconhecidos entre eles.

Os algoritmos capazes de encontrar tais relacionamentos são chamados de algoritmos de regras de associação e trabalham com a extração de conjuntos de atributos frequentes inseridos em um conjunto maior. Esses algoritmos variam bastante em relação a geração do universo de subconjuntos e em como os conjuntos de atributos escolhidos são suportados durante a geração das regras de associação.

A regra de associação tem o formato A $\longrightarrow$ B, onde A, chamado de antecedente, e B, chamado de consequente, são conjuntos de itens ou transações, e a regra pode ser lida como: o atributo A frequentemente implica no atributo B (AGRAWAL; IMIELINSKI; SWAMI, 1993). Para avaliar as regras geradas são utilizadas algumas medidas de interesse, as mais utilizadas são suporte e confiança também estabelecidas no trabalho de Agrawal, Imielinski e Swami (1993). Os autores Geng e Hamilton (2006) realizaram uma pesquisa envolvendo outras métricas e sugeriram estratégias para selecionar medidas adequadas para determinados domínios e exigências. Neste trabalho utilizaremos as seguintes medidas:

- Suporte (support): $P(AB)$. O suporte de uma regra é definido como sendo a fração de itens I que satisfazem o conjunto A e B da regra. Se o suporte não é grande o suficiente, isso significa que a regra não é digna de consideração ou que é simplesmente preterida e pode ser considerado mais tarde;
- Confiança (confidence): $P(A/B)$. É uma medida da força de suporte as regras e corresponde a significância estatística. A probabilidade de encontrar B da regra nas transações sobre a condição que essas transações também contenham A;
- Interesse (lift): $P(B|A)/P(B)$ ou $P(AB)/P(A) * P(B)$. Utilizada para encontrar dependências, ela indica o quanto mais frequente torna-se B quando A ocorre. Varia entre 0 e ∞.

Encontrar conjuntos de itens frequentes, com frequência maior ou igual à especificada pelo usuário como sendo o suporte mínimo, não é trivial, devido à explosão combinatória ocorrida ao gerar os subconjuntos de itens. Mas, uma vez que os conjuntos de itens frequentes são obtidos, é muito simples gerar regras de associação com confiança maior ou igual à especificada pelo usuário como sendo o valor mínimo (WU et al., 2008).

Nesse contexto o algoritmo Apriori, introduzido por Agrawal, Srikant et al. (1994) e apontado pela IEEE na *International Conference on Data Mining* (ICDM) (WU et al., 2008) como o mais promissor algoritmo de geração de regras de associação e umas das abordagens mais populares em mineração de dados, é um algoritmo seminal para encontrar conjuntos de itens frequentes.

Empregando busca em profundidade o algoritmo é capaz de gerar conjuntos de itens candidatos (reconhecidos como o padrão) com k elementos a partir de conjuntos de itens de $k-1$ elementos. A varredura só termina no último elemento da base de dados e padrões não frequentes são descartados.

Neste trabalho foi utilizada uma implementação do algoritmo Apriori com o uso da ferramenta *Waikato Environment for Knowledge Analysis* (Weka) (HALL et al., 2009). Essa versão do Apriori iterativamente reduz o suporte mínimo, até encontrar o número necessário de regras com a confiança mínima passada por parâmetro pelo usuário. Essa facilidade de parametrização e o fato de contar com todas as métricas de avaliação de regras de associação mencionadas anteriormente motivaram a adoção do Weka ao invés de uma nova implementação do algoritmo para extração das regras.

4.1 Pré-Processamento

Os dados somente são considerados de qualidade, se satisfizerem os requisitos do uso pretendido. Há muitos fatores que compõem a qualidade dos dados, incluindo a precisão, integridade, consistência, atualidade, credibilidade e interpretabilidade (HAN; KAMBER; PEI, 2006).

Para garantir essas medidas de qualidade, o pré-processamento dos dados foi realizado nas seguintes etapas: (a) integração dos dados, (b) limpeza dos dados, (c) redução dos dados, (d) transformação dos dados.

(a) **Integração dos dados**

As amostras de ictioplâncton coletadas no Reservatório da Usina Hidrelétrica do Lajeado e sua área de influência, como mencionado anteriormente, foram registradas em duas bases de dados diferentes. Para ajudar a reduzir redundâncias e inconsistências no conjunto de dados final foi realizada uma integração segura, a qual utilizou o código da amostra como chave de ligação entre planilhas do Excel. O conjunto final de dados foi criado em um novo arquivo do tipo CSV (*comma separated values*). Dados redundantes foram agrupados ou eliminados dependendo do valor em relação à amostra, evitando inconsistências no conjunto como um todo.

(b) **Limpeza dos dados**

Nessa etapa foram realizadas algumas rotinas para tentar assegurar a qualidade dos dados:

- Os dados faltantes foram substituídos por uma constante global assinalada por "?". Dessa forma, o algoritmo poderia lidar com as lacunas de forma a não influenciar os resultados;
- Alguns valores fora do padrão (outliers), como valores de "9999" para temperatura da água, foram retirados. Este tipo de dado foi considerado como faltante, sendo identificado pelo símbolo "?";

- Dados inconsistentes como a entrada "i" para o atributo nebulosidade, que deveria receber apenas valores numéricos por padrão, também foram assinalados como faltantes.

(c) **Redução dos dados**

O conjunto final de atributos foi reduzido do conjunto original realizando uma redução da dimensão, onde as propriedades irrelevantes, fracamente relevantes ou redundantes puderam ser detectadas e removidas.

Para tal tarefa foi empregado o algoritmo *CfsSubsetEval* que avalia o valor de um subconjunto de atributos, considerando a capacidade preditiva individual de cada recurso, juntamente com o grau de redundância entre eles. Os subconjuntos de características altamente correlacionados com a classe e com baixa intercorrelação são preferencialmente selecionados (HALL, 1999).

Para este trabalho, a combinação entre *BestFirst* (Método de busca) e *CfsSubsetEval* (atributo avaliador) é tão eficiente quanto as melhores técnicas de seleção de variáveis, como: algoritmo genético ou algoritmo *Simulated Annealing*; tendo a vantagem de ser ainda mais rápido (TETKO et al., 2006).

Para avaliar os atributos comparamos os valores utilizando a heurística mérito de cada relação formalizada pela equação (1) (GHISELLI, 1964).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

A fórmula final de mérito utiliza a correlação de Pearson entre uma variável composta (soma ou média) e uma variável de destino (a classe em questão) (HALL, 2000).

O algoritmo *CfsSubsetEval* do Weka foi executado com o conjunto inicial de dados como entrada. Dos 33 atributos iniciais, a base foi reduzida para 14. Após avaliação dos especialistas em ictiofauna, este número foi ainda mais reduzido, chegando a um total de 10 atributos tidos como essenciais para a modelagem do problema. A relação dos atributos com as respectivas pontuações de mérito está apresentada na Tabela 1.

(d) **Transformação dos dados**

- Em valores numéricos o separador decimal "," foi substituído por "." fazendo com que fossem interpretados corretamente pelo Weka;
- As datas foram formatadas no padrão "dd/mm/yyyy" uniformizando o tipo do campo;
- Algumas variáveis numéricas foram discretizadas, através do mapeamento de intervalos de valores em rótulos, como exibido na Tabela 2. Desta forma, pudemos aplicar o algoritmo Apriori que trabalha com atributos nominais.

Tabela 1 – Seleção de atributos feita com o algoritmo *CfsSubsetEval* do Weka. O atributo "estagio" refere-se ao estágio das larvas de peixes.

Atributo avaliado	Mérito	Atributos selecionados
familia	0.692	ordem, especie
ordem	0.595	smf, estagio, familia
condutividadeAgua	0.557	local, ph
especie	0.543	local, temperaturaAgua, familia
ph	0.463	margem, oxigenioDissolvido, condutividadeAgua, ordem
oxigenioDissolvido	0.445	local, margem, temperaturaAr, ph
profundidade	0.319	local, transparencia
transparencia	0.314	temperaturaAr, profundidade, oxigenioDissolvido, ph
estagio	0.268	ordem, familia
local	0.116	margem,oxigenioDissolvido,ph,condutividadeAgua, família,especie
temperaturaAgua	0.096	local, temperaturaAr
temperaturaAr	0.094	temperaturaAgua, oxigenioDissolvido, transparencia
margem	0.088	local, oxigenioDissolvido, ph, condutividadeAgua
smf	0.01	local, margem, ph, condutividadeAgua, especie

Tabela 2 – Atributos discretizados

Atributo	Valores obtidos na coleta	Valores equivalentes
Temperatura do ar	tempAr >10 e <= 15	1
	tempAr >15 e <= 20	2
	tempAr >20 e <= 25	3
	tempAr >25 e <= 30	4
	tempAr >30 e <= 35	
	tempAr >35 e <= 40	6
Temperatura da água	tempAgua >10 e <= 15	1
	tempAgua >15 e <= 20	2
	tempAgua >20 e <= 25	3
	tempAgua >25 e <= 30	4
	tempAgua >30 e <= 35	5
	tempAgua >35 e <= 40	6
Profundidade	profun >0 e <= 1	1
	profun >1 e <= 2	2
	profun >2 e <= 3	3
	profun >3 e <= 4	4
	profun >4 e <= 5	5
	profun >5 e <= 6	6
	profun >6	7
Oxigênio dissolvido	oxig >0 e <= 1	1
	oxig >1 e <= 2	2
	oxig >2 e <= 3	3
	oxig >3 e <= 4	4
	oxig >4 e <= 5	5
	oxig >5 e <= 6	6
	oxig >6 e <= 7	7
	oxig >7 e <= 8	8
	oxig >8 e <= 9	9
	oxig >9 e <= 10	10
	oxig >10 e <= 11	11
	oxig >11 e <= 12	12
	oxig >12	13
Transparência da água	transp >0 e <= 0.5	1
	transp >0.5 e <= 1	2
	transp >1 e <= 1.5	3
	transp >1.5 e <= 2	4
	transp >2 e <= 2.5	5
	transp >2.5 e <= 3	6
	transp >3	7
PH	ph <= 4.0	1
	ph >4.0 e <= 6.0	2
	ph >6.0 e <= 7.0	3
	ph >7.0 e <= 9.0	4
	ph >9.0	5
Condutividade da água	condut <= a 20	1
	condut >20 e <= 40	2
	condut >40 e <= 60	3
	condut >60 e <= 80	4
	condut >80 e <= 100	5
	condut >100 e <= 120	6
	condut >120	7

Após a finalização das etapas de pré-processamento, os dados foram reunidos em um único arquivo de tamanho reduzido contendo 9 atributos e 4913 instâncias, dos 33 atributos e 5333 instâncias originais preparados para o início dos testes que serão descritos na próxima seção.

5 Resultados

Para os experimentos, foram utilizados valores de parâmetros variados para o algoritmo Apriori objetivando encontrar as melhores regras envolvendo o atributo estágio (estágio das larvas de peixes), biótico, com o restante dos atributos abióticos.

A cobertura de uma regra de associação é tida como o número de instâncias para as quais a regra prevê corretamente, chamada de suporte. Sua precisão, chamada de confiança, é o número de instâncias que a regra prevê corretamente, expressa como uma percentagem de todas as instâncias a que se aplica (NGO, 2011).

Os valores utilizados para os parâmetros de suporte foram mantidos em 0.1 para suporte mínimo e 1.0 para suporte máximo, fazendo com que as regras fossem geradas livremente. Os valores dos parâmetros confiança e interesse foram alterados de 0.8 a 1.0 e de 1.5 a 3.5,respectivamente. No total, 979 regras de associação foram geradas, conforme o gráfico da Figura 2.

A quantidade de regras geradas com o valor da confiança acima de 0.9 e suporte entre 0.2 e 0.3 é bastante alta e visível no gráfico de dispersão da Figura 2.

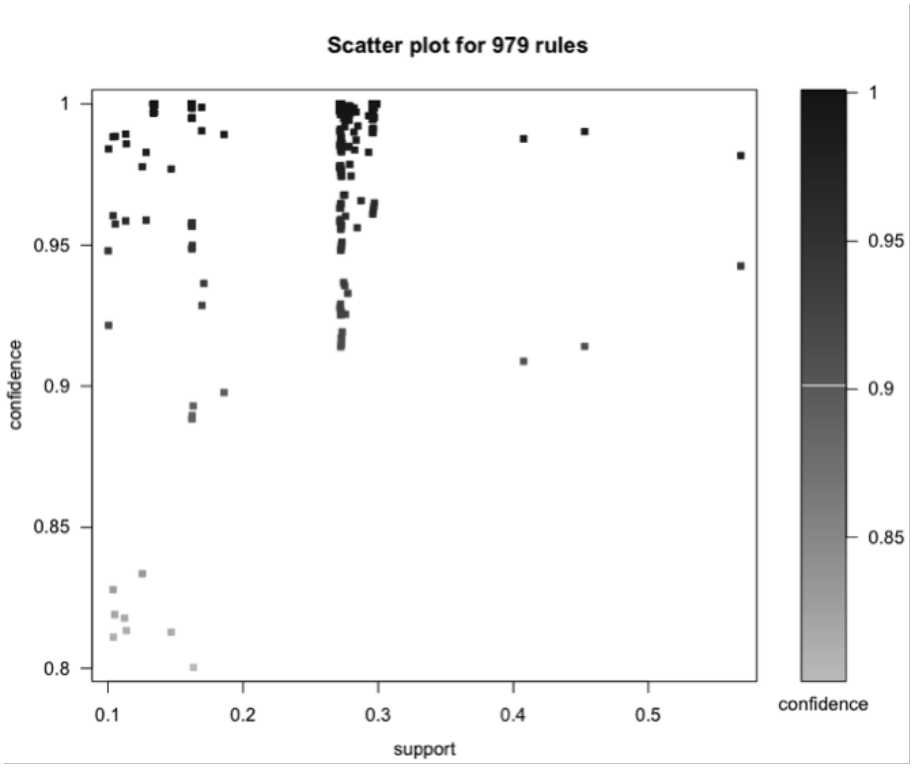

Figura 2 – Gráfico da relação entre a confiança e o suporte para as 979 regras de associação geradas e seus limites.

Para a métrica de interesse, a quantidade de regras geradas com um valor mais alto

entre 3 e 3.5, concentram-se entre o suporte 0.2 e 0.3. Regras com um valor de interesse um pouco mais baixo entre 1.5 e 2.5 tem o suporte com uma faixa de abrangência maior de 0.1 a 0.3, com raras exceções, conforme pode ser visto no gráfico da Figura 3.

A análise da confiança dos resultados das regras geradas com a métrica de interesse mostrou que as regras que têm um alto valor de confiança se encontram com a medida de suporte entre 0.2 e 0.3.

A partir dos resultados obtidos, a mineração foi refinada considerando que a quantidade de regras encontradas, ainda continha regras sem valor semântico, de acordo com a análise de especialistas em ictiofauna.

As dez melhores regras, apresentadas na Tabela 3 podem ser assim interpretadas da esquerda para a direita. O valor anterior ao símbolo "==>" indica o suporte da regra, isto é, o número de itens abrangidos pela(s) sua(s) premissa(s). O valor que aparece após o atributo consequente consiste no número de itens para os quais o consequente da regra é válido. Entre parênteses, há o valor da confiança da regra. Desta forma, conforme descrito na regra 1, podemos lê-la da seguinte forma: "*se transparência=6 e ph=5 então profundidade=7*".

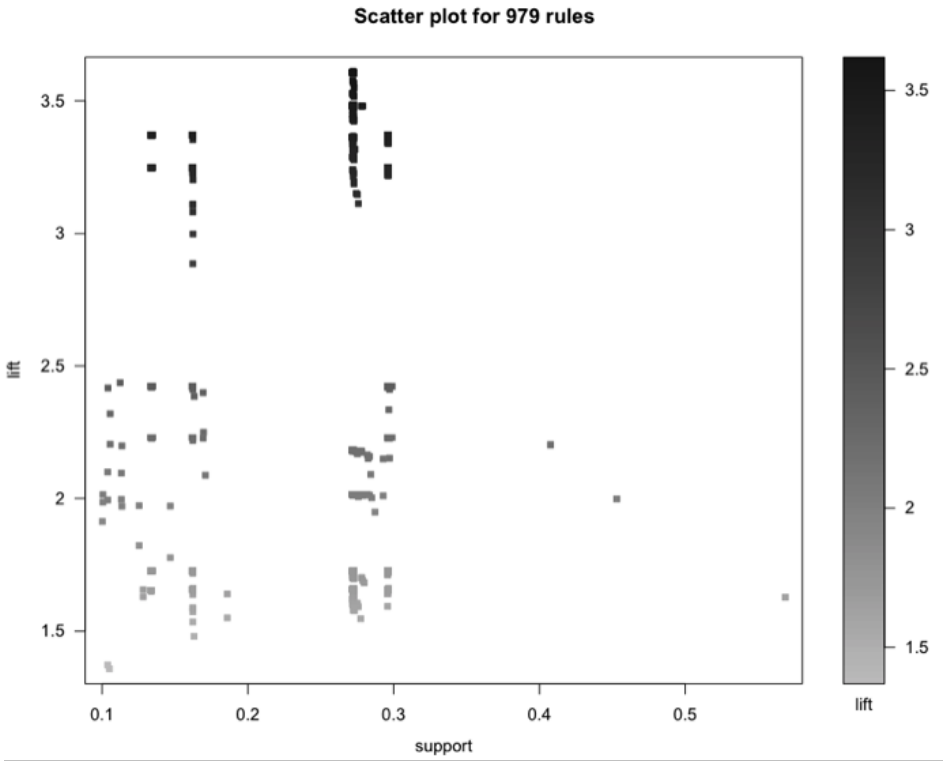

Figura 3 – Gráfico da quantidade de regras e a relação com os parâmetros interesse e o suporte.

Tabela 3 – Resultados das 10 melhores regras geradas.

1. transparencia=6 ph=5 1458 ==> profundidade=7 1458 conf:(1)
2. oxigenioDissolvido=12 transparencia=6 1456 ==> profundidade=7 1456 conf:(1)
3. oxigenioDissolvido=12 ph=5 1456 ==> profundidade=7 1456 conf:(1)
4. oxigenioDissolvido=12 ph=5 1456 ==> transparencia=6 1456 conf:(1)
5. oxigenioDissolvido=12 transparencia=6 1456 ==> ph=5 1456 conf:(1)
6. oxigenioDissolvido=12 transparencia=6 ph=5 1456 ==> profundidade=7 1456 conf:(1)
7. profundidade=7 oxigenioDissolvido=12 ph=5 1456 ==> transparencia=6 1456 conf:(1)
8. profundidade=7 oxigenioDissolvido=12 transparencia=6 1456 ==> ph=5 1456 conf:(1)
9. oxigenioDissolvido=12 ph=5 1456 ==> profundidade=7 transparencia=6 1456 conf:(1)
10. oxigenioDissolvido=12 transparencia=6 1456 ==> profundidade=7 ph=5 1456 conf:(1)

Podemos observar que o atributo estágio ainda não foi encontrado entre as 10 primeiras regras com valor de confiança igual a 1 (um). Apresentamos a Tabela 4, abaixo, que abrange o atributo estágio para as regras 17 e 22, com um valor de confiança bastante alto, de 0.99 e 0.96.

Tabela 4 – Regras que envolvem o atributo estágio.

17. transparencia=6 estagio=pre 524 ==> profundidade=7 518 conf:(0.99)
22. profundidade=7 estagio=pre 541 ==> transparencia=6 518 conf:(0.96)

As regras descobertas, apresentadas na Tabela 4, estabelecem uma relação com o objetivo inicial da pesquisa que era responder a questão da existência de alguma relação entre os fatores abióticos com os bióticos, no caso, o estágio das larvas. As duas regras foram validadas pelos especialistas em ictiofauna como importantes para o entendimento do processo da desova dos peixes.

6 Considerações Finais

As regras de associação encontradas neste trabalho com o algoritmo Apriori estão em concordância com a realidade da ictiofauna encontrada nos locais de amostragem, sendo validado semanticamente por um especialista. Esses mesmos dados utilizados já haviam sidos analisados utilizando métodos estatísticos. Porém, sem sucesso na busca de uma relação entre os meios bióticos e abióticos. A mineração de dados utilizada como análise exploratória dos dados, e, neste caso, usando regras de associação, proporcionou a descoberta de conhecimento singular para o grupo de pesquisa em ictiofauna da Amazônia Legal.

A título de exemplo do conhecimento adquirido apresentamos o caso da regra de número 17 onde o valor da transparência é 6, o que significa que há pouca transparência da água no local de coleta, e a profundidade possui valor 7, valor que indica uma das maiores profundidades encontradas ao longo dos pontos de coleta, nesse local havia presença de larvas na fase de pré-flexão, comprovando através da técnica de Mineração de Dados uma relação entre fatores abióticos e os bióticos.

Durante a pesquisa verificamos que há muitos erros na entrada dos dados que são resultantes da falha na alimentação de dados, com dados inconsistentes, e/ou *software* com deficiência para execução das tarefas de registros. Para o desenvolvimento de trabalhos futuros, e uso de técnicas de mineração de dados em outras bases de ictiofauna, estamos desenvolvendo *software* específico para a coleta dos dados provenientes das amostragens.

bem como rotinas de visualização da informação e suporte a decisão, para maximizar os resultados.

Referências

AGOSTINHO, A. A. et al. Fish ladder of lajeado dam: migrations on one-way routes? *Neotropical Ichthyology*, SciELO Brasil, v. 5, n. 2, p. 121–130, 2007. Citado na página 2.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Database mining: A performance perspective. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 5, n. 6, p. 914–925, 1993. Citado na página 5.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499. Citado 2 vezes nas páginas 2 e 6.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, Springer, v. 36, n. 1-2, p. 105–139, 1999. Citado na página 2.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 2.

GENG, L.; HAMILTON, H. J. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, ACM, v. 38, n. 3, p. 9, 2006. Citado na página 5.

GHISELLI, E. E. *Theory of psychological measurement*. [S.l.]: McGraw-Hill New York, 1964. Citado na página 7.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado na página 6.

HALL, M. A. *Correlation-based feature selection for machine learning*. Tese (Doutorado) — The University of Waikato, 1999. Citado na página 7.

HALL, M. A. Feature Selection for Discrete and Numeric Class Machine Learning. In: *Machine Learning. Proc. Seventeenth International conference on Machine Learning*. [s.n.], 2000. p. 359–366. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.6025>. Citado na página 7.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. [S.l.]: Morgan kaufmann, 2006. Citado na página 6.

HASTIE, T. et al. *The elements of statistical learning*. [S.l.]: Springer, 2009. Citado na página 2.

HOCHACHKA, W. M. et al. Data-mining discovery of pattern and process in ecological systems. *The Journal of Wildlife Management*, Wiley Online Library, v. 71, n. 7, p. 2427–2437, 2007. Citado na página 2.

MME. *Plano Decenal de Expansão de Energia 2021*. 2012. Citado na página 3.

NAKATANI, K. *Ovos e larvas de peixes de água doce: desenvolvimento e manual de identificação*. [S.l.]: Eletrobrás; Uem, 2001. il., 2001. Citado 3 vezes nas páginas 3, 4 e 5.

NGO, T. Data mining: Practical machine learning tools and technique, third edition by ian h. witten, eibe frank, mark a. hell. *SIGSOFT Softw. Eng. Notes*, ACM, New York, NY, USA, v. 36, n. 5, p. 51–52, set. 2011. ISSN 0163-5948. Disponível em: <http://doi.acm.org/10.1145/2020976.2021004>. Citado na página 9.

TETKO, I. V. et al. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. *Journal of chemical information and modeling*, ACS Publications, v. 46, n. 2, p. 808–819, 2006. Citado na página 7.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, v. 14, n. 1, p. 1–37, 2008. Citado 2 vezes nas páginas 5 e 6.

5 Explorando dados sociais para entender o trabalho infantil

Diego C. Rodrigues David N. Prata Michel A. Silva

2013

Abstract

Child labor is anissue of utmost importance to the world. According to the ILO (International Labour Organisation) nearby 218 million children between 5 and 17 work in the world, of which 50% holds hazardous work, The question rises up on how to locate and understand the factors about those families that generates the indices of child labor, and which properties are important to analyze. By the use of data mining techniques to discover valid patterns among Brazilian social data bases, we evaluate child labor in the federated state of Tocantins. This work has the purpose towards uncovering the deterministic factors for the practice of child labor and their relationships among financial, educational, cultural and social indicators; generating information that was not aware provided by data bases feeders, being hidden among records.

Key-words: data mining. social data. child labor. welfare.

Resumo

O trabalho infantil é uma questão de extrema importância para o mundo. De acordo ao OIT (Organização Internacional do Trabalho) aproximadamente 218 milhões de crianças entre 5 e 17 anos trabalham no mundo, das quais 50% têm trabalhos de risco. Através desta informação, levanta-se questionamentos sobre como localizar e entender os fatores sobre as famílias que compõem os índices de trabalho infantil, e que propriedades são importantes para analisar estes casos. Com o uso de técnicas de mineração de dados para descobrir padrões válidos sobre bases de dados sociais brasileiras, foram avaliados dados de trabalho infantil no Estado do Tocantins, localizado na região Norte do Brasil. Este trabalho tem o objetivo de detectar fatores que são determinísticos para a prática do trabalho infantil e suas relações com indicadores financeiros, educacionais, culturais e sociais, gerando informações que antes não estavam explícitas aos responsáveis pela base de dados.

Palavras-chaves: mineração de dados. dados sociais. trabalho infantil. bem-estar.

1 Introdução

O Programa de Assistência Social Cadastro Único (CadÚnico), do Governo Federal do Brasil, é um instrumento para identificar e classificar famílias de baixa renda considerando características de renda de até dois salários mínimos do país por pessoa ou renda mensal familiar de até três salários mínimos.

A base de dados do Cadastro Único permite compreender a realidade socioeconômica destas famílias, trazendo informações sobre o núcleo familiar como, por exemplo, as características de suas residências, as formas de acesso a serviços públicos essenciais e, também, as informações sobre cada um dos componentes da família.

O Governo Federal tem o objetivo de consolidar os dados coletados pelo CadÚnico, através de sistemas computacionais, para disseminar conhecimento sobre famílias de baixa renda para órgãos do poder público. Baseado nestas informações, o poder público pode ser considerado hábil para formular e implementar políticas específicas para contribuir com a redução da vulnerabilidade social a que estas famílias estão expostas. Atualmente, o programa tem mais de 21 milhões de famílias registradas.

O CadÚnico é coordenado pelo Ministério do Desenvolvimento Social e Combate à fome (MDS) e é utilizado para selecionar beneficiários de programas sociais do Governo Federal, como por exemplo, o programa Bolsa Família. Pode também ser utilizado pelos Estados e governos locais para obterem diagnósticos de registros socioeconômicos das famílias, permitindo o desenvolvimento de políticas sociais regionais.

Aplicar técnicas de mineração de dados para descobrir padrões e conhecimentos válidos não é uma tarefa trivial devido à grande quantidade de dados e atributos disponíveis no CadÚnico. Com o objetivo de analisar os dados do CadÚnico, foram aplicadas técnicas de mineração para descobrir padrões relacionados ao trabalho infantil no Estado do Tocantins. A base de dados foi dividida em dois conjuntos, combinando dados de famílias com 168 atributos e dados de pessoas com 214 atributos, totalizando um conjunto de 892.422 registros para qualificar diversificadas informações.

De acordo ao IBGE (Instituto Brasileiro de Geografia e Estatística), nos registros do censo 2010, a região Norte do Brasil está em primeiro lugar em índices de trabalho infantil (IBGE, 2010). A partir disto, algumas questões necessitam ser identificadas, como a necessidade de evidenciar a situação destas famílias apontadas pelo IBGE com o intuito de se certificar de que as ações do governo têm capacidade de atingir e apoiar estas pessoas.

No entanto, um fato problemático surge: como saber se esta família realmente tem todas as características relevantes para ter este tipo de atenção do governo? Apesar da quantidade de dados coletadas trimestralmente pelo Governo Federal, pode haver falhas nos registros ou registro com omissões.

De acordo ao censo 2010 (IBGE), o Estado do Tocantins tem uma estimativa de 40 mil domicílios com índices de trabalho infantil (IBGE, 2010). Porém, na verdade, o CadÚnico apresenta uma quantidade de 5.250 famílias com este tipo de trabalho. Esta inconsistência induz o governo a realizar uma busca para localizar e esclarecer quais famílias estão inscritas inadequadamente nos programas sociais e quais não estão inscritas, mas deveriam estar, e realizar as devidas correções. Sobre este tema, as atenções estão voltadas para localizar e conhecer os fatores que estão relacionados às famílias que geram índices de trabalho infantil, e quais atributos são realmente importantes.

Este trabalho tem o propósito de responder as seguintes questões: "*Quais são os*

fatores determinístico para a prática do trabalho infantil?"e "*Quais são os reais relacionamentos entre os indicadores financeiros, educacionais, culturais e sociais?*".

De fato, existe uma forte tendência em associar o trabalho infantil à condição financeira familiar da criança, mas este fator é realmente um indicador determinístico? E, que tal considerar outros indicadores?

2 Organização dos dados

O CadÚnico é mantido há mais de 15 anos pelo governo brasileiro. Esta base de dados tem a função de registrar todas as famílias brasileiras que estão sob risco social. Armazena um conjunto de dados de famílias e seus membros, criando um poderoso conjunto de dados com grande potencial de informações.

A base de dados está dividida em dois grupos, sendo eles família e indivíduos. No conjunto de dados sobre as famílias estão todas informações acerca do tipo de habitação, renda familiar e informações sociais, e sobre as condições financeiras da família. Os dados sobre os indivíduos pertencentes às famílias são suas características e dados pessoais, como informações escolares, sociais e financeiras referente a cada indivíduo. Assim, o governo gera um registro completo das famílias e seus membros.

3 Descoberta do conhecimento

Mineração de dados é parte de um processo de pesquisa maior conhecido como Descoberta de Conhecimento em Bases de Dados, do inglês Knowledge Discovery in Databases (KDD), definido como a exploração e análise, automática ou semiautomática, de grandes quantidades de dados com o intuito de descobrir padrões e regras relevantes (LINOFF; BERRY, 2011).

As técnicas de mineração de dados estão em crescente popularidade como ferramenta para o KDD na busca por informações desconhecidas, necessárias para o processo de tomada de decisão. No entanto, esta abordagem é de difícil aplicação por causa de suas habilidades interdisciplinares para combinar diferentes métodos e técnicas, tais como banco de dados, métodos estatísticos, redes neurais, algoritmos genéticos, aprendizado de máquina, processamento de linguagem natural e outras áreas de estudo.

Os dados a serem trabalhados nem sempre estão em perfeitas condições para iniciar o processo de mineração. Dados têm normalmente muitas fontes, podem estar incompletos ou conter ruídos. Estes dados sem tratamento necessitam de um passo importante conhecido por pré-processamento, incluindo atividades como limpeza, integração, seleção e transformação dos dados (AGARWAL et al., 1996).

Após a finalização do pré-processamento, os dados podem ser carregados no software de mineração de dados, como o Weka (*Waikato Environment for Knowledge Analysis*) (HALL et al., 2009). O Weka inclui uma série de algoritmos para formatação de dados, algoritmos para aprendizado de máquina e validação de resultados, escrito na linguagem de programação Java (ORACLE, 2013), com seu código aberto disponível na internet.

Dentre as técnicas de mineração de dados disponíveis, a tarefa de classificação pode ser considerada uma das mais adequadas técnicas para fins de estudos científicos.

O objetivo da análise dos dados foi classificar indicadores sociais com o intuito de descobrir padrões que não estavam explícitos sobre os dados relacionados ao trabalho

infantil.

4 Árvores de decisão

A tarefa de classificação utiliza um tipo especial de inteligência artificial conhecida como árvore de decisão. Baseado em registros de conjuntos de treinamento (o conjunto de dados usados para treinar o algoritmo), uma árvore é gerada. O projeto de uma árvore de decisão nos permitiu classificar dados por amostragem, descartando a necessidade de testar todos os campos disponíveis, o que torna este tipo de classificação um dos melhores métodos em relação à performance e um dos algoritmos de mineração de dados mais utilizados (WU et al., 2008).

Em uma árvore de decisão, o conhecimento é representado por nós e ramos organizados de forma hierárquica. Cada nó pai permite o acesso aos seus nós filhos. Assim, da raiz às folhas da árvore as configurações do sistema são modeladas, assim como o comportamento associado.

A qualidade da geração das árvores pelo sistema foi analisada por um grupo de especialistas em assistência social, onde foi avaliada a classificação dos indicadores quanto a sua relevância em relação à possíveis fatores que levam ao trabalho infantil.

5 Metodologia

O primeiro estágio do trabalho consistiu em examinar algoritmos de mineração de dados e escolher o mais adequado para encontrar padrões entre os atributos em análise. O algoritmo *J48* foi apontado pelo IEEE na *International Conference on Data Mining* (ICDM) (HALL, 1999) como o mais promissor algoritmo para geração de árvores de decisão e uma das mais populares abordagens em mineração de dados, o algoritmo 1 detalha o funcionamento do *J48*.

Algoritmo 1: Pseudo código do algoritmo J48

```
Input: um conjunto de dados D
begin
    Árvore={};
    if D é "puro" OU existe outro critério de parada then
        encerrar;
    foreach atributo a ∈ D do
        Calcular ganho de informação;
    end
    a_melhor = Melhor atributo de acordo com o calculo do ganho de informação;
    Árvore = Cria um nó baseado no a_melhor;
    D_v = Divide o subconjunto de D baseado no a_melhor;
    foreach D_v do
        Árvore_v = J48(D_v);
        Fixe a Árvore_v no galho obtido no passo anterior;
    end
end
return [Árvore]
```

Neste trabalho foi utilizada uma implementação do algoritmo de geração de árvore de decisão através da ferramenta Weka (HALL et al., 2009). Esta versão do algoritmo conhecido como *J48* é uma implementação do algoritmo *C4.5* (QUINLAN, 1993) para o Weka. O algoritmo utiliza a técnica do método guloso (*greedy method*) para realizar a classificação de árvores de decisão e a técnica de poda para redução de erros. Por ser uma ferramenta de fácil utilização e oferecer uma grande quantidade de recursos, o Weka foi adotado para este trabalho ao invés de uma nova implementação do algoritmo de geração de árvores.

5.1 Pré-Processamento

Os dados somente são considerados de qualidade, se satisfizerem os requisitos do uso pretendido. Há muitos fatores que compõem a qualidade dos dados, incluindo a precisão, integridade, consistência, atualidade, credibilidade e interpretabilidade (TETKO et al., 2006). Para assegurar essas medidas de qualidade no pré- processamento, alguns passos são necessários, como segue:

(a) **Integração dos dados**

A base de dados do CadÚnico foi armazenada em diferentes tabelas. Para auxiliar na redução de redundâncias e inconsistências no conjunto de dados de trabalho foi realizada uma integração segura, a qual utilizou o código da amostra, chave única de ligação entre a tabela que continha informações das pessoas e das suas respectivas famílias. O conjunto final foi então gerado em um arquivo do tipo CSV (*comma-separated values*). Dados redundantes foram agrupados ou eliminados dependendo do valor em relação à amostra, evitando inconsistências no conjunto de dados.

(b) **Limpeza dos dados**

Neste estágio algumas rotinas foram realizadas na tentativa de assegurar a qualidade dos dados, onde dados faltantes foram substituídos pela constante global "?". Desta forma o algoritmo poderia lidar com as lacunas de forma a não influenciar o resultado final.

(c) **Redução dos dados**

Após a limpeza dos dados, o conjunto final de atributos foi reduzido, a partir do original, realizando uma redução da dimensão onde atributos fracamente relevantes ou dados redundantes puderam ser detectados e eliminados.

Nesta tarefa foi empregado o algoritmo *CfsSubsetEval* para avaliar o valor de um subconjunto de atributos, considerando a capacidade de previsão de cada característica, assim como o grau de redundância entre eles. Os subconjuntos de características altamente correlacionados com a classe e com baixa intercorrelação são preferencialmente selecionados (HALL, 1999).

Para este trabalho, a combinação entre *BestFirst* (método de busca) e *CfsSubsetEval* (atributo avaliador) é tão eficiente quanto as técnicas de seleção de variáveis, como algoritmo genético e o algoritmo *Simulated Annealing*, além de ser mais rápido (TETKO et al., 2006).

Para avaliar os atributos, foram comparados valores utilizando-se da heurística mérito de cada relação formalizada pela Equação 1 (GHISELLI, 1964).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

A fórmula final de mérito utiliza a correlação de Pearson entre uma variável composta (soma ou média) e uma variável de destino (a classe em questão) (HALL, 1999).

O algoritmo *CfsSubsetEval* do Weka foi executado com o conjunto inicial de dados como entrada. Dos 92 atributos iniciais, a base foi reduzida em 35. Após a avaliação dos especialistas em assistência social este número foi reduzido novamente e totalizou 5 atributos considerados essenciais para a modelagem do problema. A lista de atributos com suas respectivas pontuações de mérito está apresentada na Tabela 1.

Tabela 1 – Seleção de atributos feita com o algoritmo *CfsSubsetEval* do Weka.

Atributo avaliado	Mérito	Atributos selecionados
sexo da criança	0.01	Índice de Trabalho Infantil
Pessoa Recebe Beneficio do Governo (PBF)	0.084	Frequência Escolar, Renda Familiar
Frequência Escolar	0.018	Pessoa Recebe Benefício do Governo (PBF), Índice de Trabalho Infantil
Renda Familiar	0.084	Pessoa Recebe Benefício do Governo (PBF)
Índice de Trabalho Infantil	0.018	Sexo da Criança, Frequência Escolar

Além disso, foram relacionados os dados de cada família com seus respectivos membros e em seguida, após a execução de algoritmos de seleção de atributos citados, os mesmos atributos selecionados pelo algoritmo foram eleitos pelos especialistas como os indicadores mais relevantes para o problema em questão.

(d) **Transformação dos Dados**

Os separadores "," dos valores numéricos decimais foram substituídos por "." para interpretação correta pelo Weka e datas foram formatadas para o padrão "dd/mm/yyyy", valores nulos foram substituídos por "?", e transformados de dados numéricos para nominal.

6 Experimentação

O experimento foi dividido em 4 etapas:

1. A primeira etapa para a fase de experimentos foi a identificação da não conformidade da qualidade dos dados especialmente pela quantidade de dados faltantes no banco de dados do CadÚnico. Esta evidência requer um refinamento dos dados que estão em 840.000 registros verificando a existência de campos em branco e valores nulos para os 5 atributos selecionados (sexo da criança, renda familiar, frequência escolar, índice de trabalho infantil, existência de assistência pelo programa Bolsa Família), com o intuito de ajustar a base de dados para ser interpretada pelo pacote do *software* Weka.

2. A segunda etapa foi de seleção e junção dos dados com o uso da linguagem *Structured Query Language* (SQL), que suporta manipulação de dados através da seleção de registros sem valores faltantes em seus campos. Este passo permitiu validar e preparar todos os dados para serem exportados para o formato de arquivo CSV (*Comma - separated values*).

3. A terceira etapa consistiu na identificação e agrupamento do mais relevante indicador de trabalho infantil. O objetivo foi criar um conjunto de dados a partir da base do CadÚnico representando apenas os registros nos quais o atributo "trabalho infantil" estava marcado, positivo ou negativamente, deixando fora do conjunto os registros que tinham este campo com valores faltantes. O resultado foi um total de 299.614 registros com este campo preenchido. Este passo permitiu avaliar o contexto de dados de forma confiável e de aplicar as técnicas de descoberta de conhecimento.

4. A quarta etapa foi validar o refinamento da base de dados do CadÚnico com os especialistas em assistência social do governo. Todos os indicadores selecionados foram apresentados para a equipe do Ministério do Desenvolvimento Social e Combate à fome para que o conjunto de dados fosse avaliado e validado por eles. Assim, assegurou-se que o conjunto de dados era confiável e podia ser utilizado na mineração de dados.

Após a conclusão da etapa de pré-processamento, o conjunto de dados foi importado pelo Weka para que os dados fossem classificados utilizando o algoritmo *J48*, gerando a árvore de decisão, Figura 1.

7 Resultados e discussões

O resultado da árvore gerada pelo algoritmo foi apresentado para o grupo de especialistas em assistência social do governo do Tocantins, que avaliaram e interpretaram as saídas em conjunto com os especialistas em mineração de dados.

O atributo utilizado como base para checar os dados relacionados com o trabalho infantil foi "Pessoa com benefício do Bolsa Família" (PBF), Programa de Assistência Social do Governo, que constatou a incidência de trabalho infantil em relação aos outros atributos (renda familiar, incidência de trabalho infantil, sexo da criança, frequência escolar). Fatores sociais, como rendimentos educacionais para a incidência de trabalho infantil, também foram analisados.

Para o caso de "Ausência de trabalho infantil" (a parte esquerda da árvore apresentada na Figura 1), é possível observar o primeiro nó SIM, que significa que a família da criança recebe assistência do programa PBF (Programa Bolsa Família). Isto quer dizer que para o conjunto de dados que compreende a base de dados do CadÚnico do Estado do Tocantins versão 2013, o trabalho infantil foi erradicado das famílias que receberam benefícios do PBF. Olhando pelo ponto de vista social, esta informação é muito importante a ser considerando o risco social do trabalho infantil para todas estas famílias que estão em situação de baixa renda. Isto é um indicativo de que a assistência governo favorece a erradicação do trabalho infantil.

O fator "Renda Familiar" foi outro campo bastante observado na árvore gerada, em especial o preenchimento deste atributo com valor de até R$ 66,50 (sessenta e seis reais e cinquenta centavos), em que não foram verificados casos de trabalho infantil. Por outro lado, pode-se observar que os nós da árvore com renda familiar entre R$ 66,50 (sessenta e seis reais e cinquenta centavos) e R$ 133,00 (cento e trinta e três reais) apresentam casos de trabalho infantil, o que induz a argumentos de que o atributo "renda familiar" pode não ser conclusivo sobre o trabalho infantil. Esta informação precisa de um estudo mais aprofundado e uma pesquisa mais extensa para que se entenda a motivação dos casos de trabalho infantil nas famílias, em relação à renda familiar.

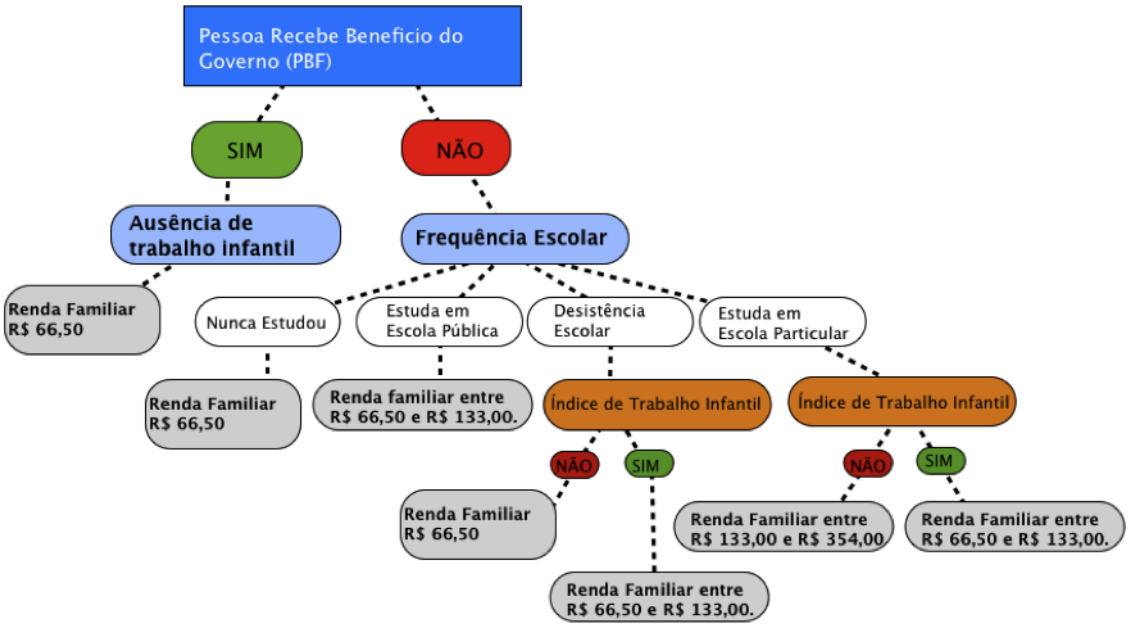

Figura 1 – Árvore de Decisão gerada pelo algoritmo *J48* do Weka

Esta evidência fornece embasamento para que o governo possa tomar melhores decisões sobre quais ações promover e que tipos de campanhas adotar contra o trabalho infantil.

O atributo "Frequência Escolar" tem quatro possíveis valores para a escolarização da criança (nunca estudou, estuda em escola pública, desistência escolar, estuda em escola particular). Pode ser notado que estes fatores escolares tiveram um grande relacionamento com ocorrências de trabalho infantil. Por exemplo, para o atributo "nunca estudou" não existe ocorrência de trabalho infantil, mas para crianças que já estudaram e que houve desistência escolar existe ocorrência de trabalho infantil.

Um ponto interessante para o atributo "Frequência Escolar" é a ocorrência do trabalho infantil para crianças que estudam em escola privada. Para este nó na árvore se pode observar que há incidência de trabalho infantil para rendas entre R$ 66,50 (sessenta e seis reais e cinquenta centavos) e R$ 133,00 (cento e trinta e três reais) e o índice foi marcado como negativo quando as rendas estão entre R$ 133,00 (cento e trinta e três reais) e R$ 354,00 (trezentos e cinquenta e quatro reais). Este resultado nos sugere considerar algumas hipóteses, como por exemplo, questionar por que a criança está em uma escola particular. Então, algumas explicações podem ser levantadas, como a justificativa da criança trabalhar e se manter em uma escola particular para ter melhor qualidade ou estas crianças recebem bolsas para estudar. Mesmo que algumas famílias não necessitem de ajuda do governo, se comparadas às famílias que se encontram em situação financeira pior, as famílias destas crianças devem estar em busca de um ensino de melhor qualidade.

Por último, os resultados e discussões sugerem que o governo deve se utilizar das técnicas de mineração de dados para tomar melhores decisões de assistência social.

8 Conclusão

Trabalho infantil é um assunto de extrema importância para o mundo. Segundo com a OIT (Organização Internacional do Trabalho) aproximadamente 218 milhões de crianças entre 5 e 17 anos trabalham em todo o mundo, dos quais 50% possuem trabalhos de risco (PAULO, 2007).

Ásia, frica e America do Sul têm as maiores taxas de trabalho infantil, no Áentanto, os países ricos e desenvolvidos não se excluem por completo deste Ácontexto.

De acordo ao UNICEF (Fundo das Nações Unidas para a Infância), a Índia possui aproximadamente 14% de crianças entre 5 e 14 anos de idade que estão envolvidos em atividades de trabalho infantil. A maioria destas crianças trabalham em casas de família e com subcontratos. A supervisão governamental não oferece atenção a estes casos e estas crianças estão sujeitas à exploração. Hoje em dia, o continente asiático corresponde à maior taxa de trabalho infantil no mundo.

Esta situação é preocupante porque estas crianças ficam expostas a qualquer tipo de trabalho em busca do sustento de sua família, distante da proteção de um lar.

No Brasil, políticas de combate ao trabalho infantil são oferecidos pelo Governo Federal. Aplicar técnicas de descoberta do conhecimento e padrões é uma forma de potencializar os resultados a serem analisados, auxiliando a identificação, localização e a entender os casos de crianças que estão na situação de risco social, especialmente no Brasil que é o 5° maior país em extensão geográfica no mundo. Mineração de dados pode permitir buscas efetivas por atividades sociais que se potencialize um risco, assim, suprindo os responsáveis com informação, conhecimento e oferecendo suporte às decisões de quais ações são necessárias para combater os fatores deste problema global.

Os resultados desta pesquisa sugerem como trabalhos futuros estudos que possam realizar a comparação de dados considerando registros de antes e depois do fornecimento de auxílios às famílias, bem como o desenvolvimento de pesquisas comparando dados sobre o trabalho infantil entre os Estados brasileiros, verificando seu comportamento.

Referências

AGARWAL, S. et al. On the computation of multidimensional aggregates. In: *VLDB*. [S.l.: s.n.], 1996. v. 96, p. 506–521. Citado na página 3.

GHISELLI, E. E. *Theory of psychological measurement*. [S.l.]: McGraw-Hill New York, 1964. Citado na página 5.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado 2 vezes nas páginas 3 e 5.

HALL, M. A. *Correlation-based feature selection for machine learning*. Tese (Doutorado) — The University of Waikato, 1999. Citado 3 vezes nas páginas 4, 5 e 6.

IBGE. *Censo do trabalho infantil no Brasil em 2010*. 2010. Disponível em: <http://censo2010.ibge.gov.br/trabalhoinfantil>. Citado na página 2.

LINOFF, G. S.; BERRY, M. J. *Data mining techniques: for marketing, sales, and customer relationship management*. [S.l.]: John Wiley & Sons, 2011. Citado na página 3.

ORACLE. *The Java Programming Language and the Java Platform*. 2013. Disponível em: <http://www.oracle.com/technetwork/topics/newtojava/downloads/index.html>. Citado na página 3.

PAULO, F. de S. *218 milhões de crianças trabalham no mundo*. 2007. Disponível em: <http://www.estadao.com.br/noticias/internacional,218-milhoes-de-criancas-trabalham-no-mundo-calcula-oit,10222,0.htm>. Citado na página 9.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Morgan kaufmann, 1993. Citado na página 5.

TETKO, I. V. et al. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. *Journal of chemical information and modeling*, ACS Publications, v. 46, n. 2, p. 808–819, 2006. Citado na página 5.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, v. 14, n. 1, p. 1–37, 2008. Citado na página 4.

6 Mineração de Dados como ferramenta de análise exploratória em Sistemas Ecológicos

Michel de A. Silva Daniela Q. Trevisan David N. Prata
Patrick Letouze Elineide E. Marques

Abstract

Data Mining techniques can be used as an ally to the traditional statistical techniques when the knowledge of a given domain is scarce and the chances are not developed to explore the data in ecological systems, serving as support for exploratory analysis.

Key-words: data mining. ecological systems. exploratory analysis. ecosystems.

Resumo

As técnicas de Mineração de Dados podem ser utilizadas como aliada às técnicas estatísticas tradicionais quando o conhecimento de um determinado domínio é escasso e as hipóteses não estão desenvolvidas para explorar os dados em sistemas ecológicos, servindo como suporte a Análise Exploratória.

Palavras-chaves: mineração de dados. sistemas ecológicos. análise exploratória. ecossistemas.

1 Introdução

Métodos estatísticos são largamente utilizados pelos ecologistas para análise das relações entre uma resposta observada e um conjunto de preditivas sobre um conjunto de dados. Esta abordagem para análise de dados é mais apropriada para teste de hipóteses em que o conhecimento acerca dos dados e do assunto é amplo, sendo que das diversas técnicas estatísticas para este tipo de análise, as mais aplicadas são testes de hipóteses, análise de variância, ANOVA, intervalos de confiança, erro-padrão, valor-p, poder estatístico e distribuições de frequência (DENNIS, 2004).

Quando o conhecimento inicial é mínimo e as hipóteses não estão claramente desenvolvidas, as análises exploratórias são mais apropriadas do que as análises confirmatórias (HOCHACHKA et al., 2007). Neste sentido, as técnicas de mineração de dados são consideradas mais poderosas, flexíveis e eficientes para análise exploratória do que

as técnicas estatísticas (BREIMAN, 1996), (BAUER; KOHAVI, 1999), (HASTIE et al., 2005).

Este trabalho pretende apresentar a Mineração de Dados como um recurso importante para realização da análise exploratória em sistemas ecológicos apresentando a distinção entre ela e a análise confirmatória. Além disso, mostrar que é possível utilizar os métodos de Mineração de Dados para gerar hipóteses a partir dos dados e servir como apoio a pesquisas futuras.

Para isto, será apresentado aqui conceitos essenciais de Mineração de Dados e sua colocação dentro do processo Descoberta do Conhecimento em Bases de Dados (KDD), os principais métodos de mineração de dados com suas devidas divisões, citando os principais algoritmos para minerar dados.

A partir disto, este trabalho mostra a distinção de duas pesquisas realizadas sobre uma mesma base de dados. O primeiro com resultados obtidos através de métodos estatísticos e realizada por especialistas no assunto, e o segundo trabalho com técnicas de Mineração de Dados, através da busca por padrões desenvolvido por mineradores não especialistas. Ao final, observa-se que as técnicas citadas são complementares para a análise exploratória de sistemas ecológicos e que a mineração de dados tem grandioso potencial nesta área para construção de modelos cada vez mais eficientes.

2 Mineração de Dados

Para auxiliar o entendimento do trabalho se faz necessária uma apresentação dos conceitos que permeiam as tecnologias utilizadas, como é o caso da Mineração de Dados, do inglês *Data Mining*.

A quantidade de dados está cada vez maior e deve continuar aumentando com a redução do custo dos equipamentos necessários para armazenamento e com a disponibilização de serviços que fazem isso de forma online através de *cloud computing* (computação nas nuvens). Esse crescimento no volume dos dados remete a uma questão que é: Como explorar esses dados e extrair informação?

A exploração de dados para extrair padrões e realizar previsões não é algo novo, segundo Witten, Frank e Hall (2011) economistas, estatísticos, analistas e engenheiros de comunicação há muito tempo trabalham com a ideia de que padrões nos dados podem ser procurados automaticamente, identificados, validados e utilizados para fins de previsão. O que é novo é o aumento vertiginoso em oportunidades para encontrar padrões em dados. O crescimento desenfreado de bases de dados nos últimos anos, os bancos de dados para tais atividades cotidianas como as escolhas dos clientes, traz a mineração de dados para a vanguarda das novas tecnologias de negócios.

Neste mesmo sentido, se torna interessante a análise de dados ecológicos através da busca de padrões para que conhecimentos sejam revelados, mesmo que não possua ainda uma hipótese clara a ser explorada.

2.1 Descoberta do Conhecimento (KDD)

Apesar de Mineração de Dados ser o termo mais conhecido, ele é na verdade parte de um processo maior denominado Descoberta do Conhecimento, do inglês *Knowledge Discovery in Databases* (KDD), que segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) é o processo não trivial de identificar padrões em dados que sejam válidos, novos, potencialmente

úteis e compreensíveis, visando melhorar o entendimento de um problema ou um processo de tomada de decisão.

A Figura 1 fornece uma visão geral sobre o processo do KDD segundo Hong, Kuo e Chi (1999) e situa o papel da Mineração de Dados no processo como um todo.

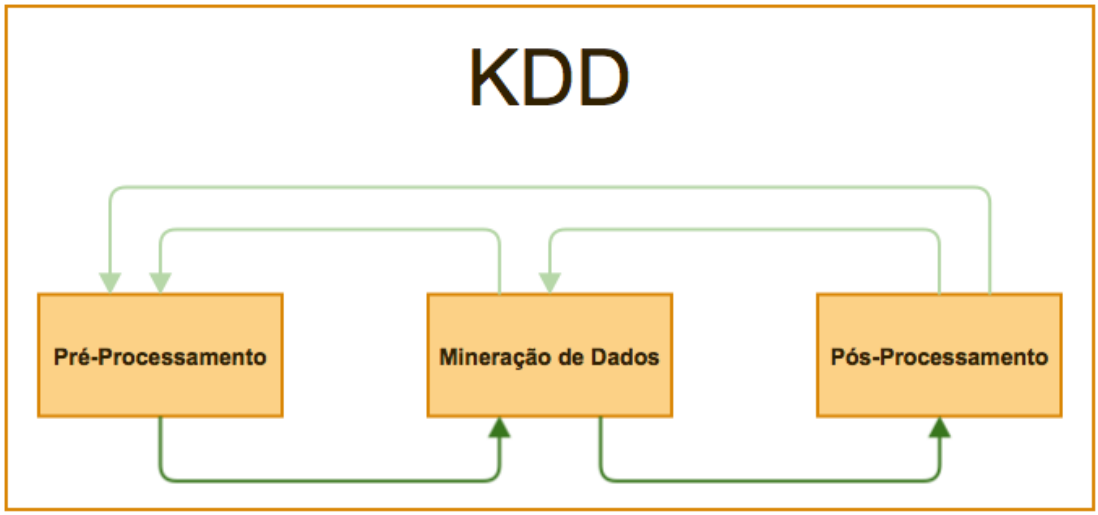

Figura 1 – Diagrama que representa visão geral das etapas do processo de KDD.

A Mineração de Dados de dados é o núcleo do processo KDD, envolve algoritmos que exploram os dados, desenvolve modelos e descobre padrões antes desconhecidos.

2.2 Objetivos

São referenciados dois grandes ramos como objetivos da Mineração de Dados: a descoberta e a verificação. Um sistema baseado em verificação tem como objetivo verificar hipóteses do usuário, enquanto um sistema de descoberta é projetado para descobrir padrões automaticamente. A descoberta de padrões se subdivide em outros dois ramos: predição e descrição. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) e Han e Kamber (2006) enquanto os algoritmos descritivos consistem na identificação de comportamentos característicos do conjunto de dados e interpretáveis por humanos, os algoritmos baseados em predição realizam previsões com base em experiências anteriores ou através de exemplos aprendidos, nesse interim as máquinas de aprendizado também são incluídas.

Mineração de dados descritiva consiste em buscas sobre enormes conjuntos de dados para descobrir a localização de estruturas inesperadas ou relacionamentos, padrões, tendências, agrupamentos (*clusters*) e valores discrepantes (*outliers*) nos dados. Por outro lado, previsão é a construção de modelos e procedimentos para a regressão, classificação, reconhecimento de padrões ou tarefas de aprendizagem de máquina, e avaliar a precisão da previsão dos modelos e procedimentos quando aplicado a novos dados.

Devido à complexidade dos bancos de dados as técnicas de mineração de dados orientadas a verificação precisam ser complementadas com técnicas orientadas a descoberta de padrões. Nada impede de que uma técnica complemente a outra e nem que os modelos construídos sejam avaliados, inclusive esse é um comportamento estimulado para obter melhores resultados.

2.3 Algoritmos populares

Existem inúmeros métodos para utilização em Mineração de Dados, por vezes o método a ser aplicado depende do tipo de dado de entrada e o tipo de informação desejada como saída. Como dito anteriormente, os métodos se dividem entre métodos de verificação e métodos de descoberta e entre os de descoberta, que são os mais interessantes do ponto de vista da análise exploratória, se dividem ainda em métodos de predição e métodos de descrição. Esta divisão está ilustrada na Figura 2.

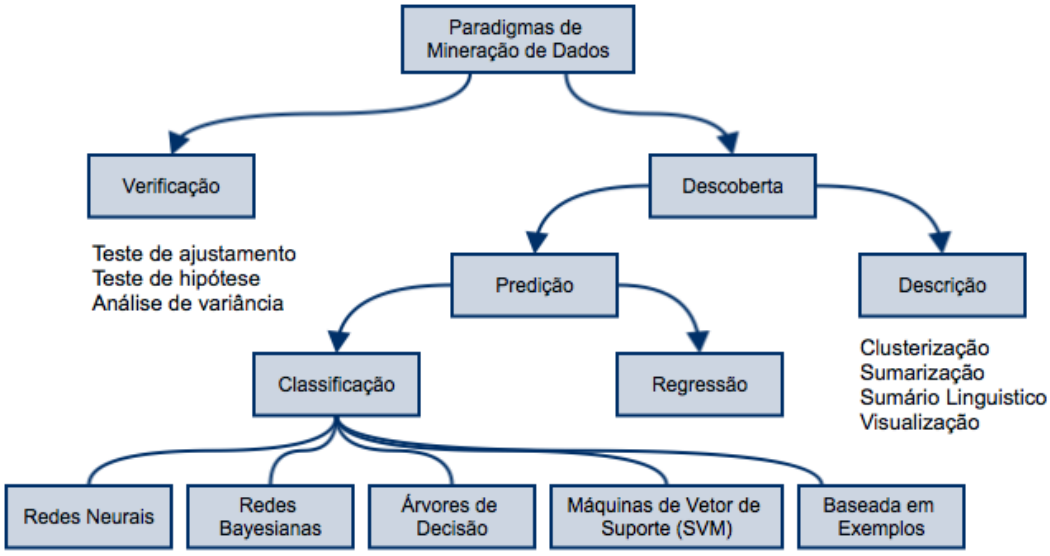

Figura 2 – Taxonomia dos métodos de Mineração de Dados segundo Maimon e Rokach (2010).

Entre as técnicas exibidas na Figura 2 alguns algoritmos têm se provado mais eficazes na tarefa de mineração. Como visto em Wu et al. (2008) os 10 algoritmos mais utilizados na Conferência Internacional IEEE em *Data Mining* (ICDM) em dezembro de 2006 foram *C4.5*, *k-Means*, *SVM*, *Apriori*, *EM*, *PageRank*, *AdaBoost*, *kNN*, *Naive Bayes*, e *CART*.

3 Analisando dados com Mineração de Dados e estatística

John W. Tukey cunhou a expressão Análise Exploratória de Dados e foi figura notável nos campos das ciências físicas e sociais. Em se tratando de análise de dados, suas contribuições incluem o algoritmo da transformada de Fourier e a Análise Exploratória de Dados. Ele deu novo impulso à estatística descritiva através da Análise Exploratória de Dados e mudou a linguagem e o paradigma dos estatísticos (TUKEY; BRILLINGER, 1986).

A Análise Exploratória de Dados fornece um quadro mais amplo sobre os métodos para estudo detalhado dos dados, antes de adaptá-los. Tem como finalidade obter a maior quantidade de informação possível, que forneça indicações sobre modelos aceitáveis que poderiam ser utilizados adiante, já na análise confirmatória.

Uma das mais notáveis diferenças é a questão do trato com o modelo de avaliação e validação. Na estatística o ajuste do modelo é mais frequentemente avaliado para determinar

se um modelo é melhor do que algum outro ou se algum modelo encaixa os dados melhor do que outro (HOCHACHKA et al., 2007).

Do ponto de vista estatístico a Mineração de Dados pode ser vista como uma análise exploratória dos dados encontrados em grandes conjuntos, realizada automaticamente. E é justamente nesse ponto que a análise confirmatória diverge da exploratória, já que ela parte da premissa que existe uma hipótese sobre quais variáveis medem quais fatores e deseja-se confirmar o grau de ajuste dos dados observados à teoria hipotetizada.

As estatísticas clássicas, que são parte do paradigma frequentista, são o tipo de estatística apresentada em aulas introdutórias (BOLKER, 2008). Para um procedimento experimental, calcula-se a probabilidade de um determinado resultado, o qual é definido como frequência média de longo prazo em que ocorre o resultado em determinada sequência de experimentos. O próximo passo é calcular o valor-p definido como a probabilidade daquele resultado ou um resultado mais extremo dada uma hipótese nula especificada. Se essa probabilidade, chamada probabilidade caudal, é pequena, então rejeita-se a hipótese nula, caso contrário, aceita-se. Mas não aceitar a hipótese alternativa se a probabilidade caudal é grande, falhou-se em rejeitar a hipótese nula (BHATTACHARYYA, 1977).

Estas metodologias são largamente aceitas dentro de pesquisas na área de ecologia e podem ser vistas em trabalhos diversos. Com o objetivo de apresentar um trabalho com estas características, foi levado em consideração o trabalho de Freitas (2010), que apresenta dentre outras técnicas de estatística, o cálculo de frequência da distribuição de estágios de ovos e larvas na área de influência da Usina Hidrelétrica do Lajeado, na bacia do Rio Tocantins, com resultados representados na Figura 3.

Quando o conhecimento prévio é insuficiente e não há o bastante para produzir um modelo paramétrico que atenda eficientemente, utiliza-se vários modelos que tentam preencher essa lacuna. Um exemplo dessa utilização é o Critério de Informação de Akaike (AIC), o qual é cada vez mais utilizado em análises no campo da ecologia. Esta medida permite comparar e classificar vários modelos concorrentes e estimar qual deles melhor se aproxima do "verdadeiro" processo subjacente ao fenômeno biológico em estudo (SYMONDS; MOUSSALLI, 2011). Nesse caso a avaliação do modelo utilizou os mesmos dados usados para construir o modelo que, diferente do paradigma da Mineração de Dados, enfatiza que para validar modelos e mensurar desempenho é necessário a criação de um conjunto de dados de teste, e a abordagem inclui ainda a validação cruzada e *bootstrap*.

No trabalho de Makler-Pick et al. (2011) utilizou-se vários métodos estatísticos como GLM do inglês *General Linear Model* que é uma generalização flexível do modelo de regressão linear e RPART do inglês *Recursive Partitioning And Regression Trees* que cria uma árvore hierárquica que é útil para captura de relações não-lineares complexas.

No entanto, mesmo com toda essa flexibilidade ao lidar com modelos pode ser insuficiente para permitir aos ecologistas extrair noções biológicas dos seus dados, se o conhecimento prévio é mínimo e as hipóteses não estão claramente desenvolvidas. Nesse iterim as análises exploratórias são úteis para gerar hipóteses sendo assim mais apropriadas que as análises confirmatórias normalmente apresentadas em publicações ecológicas.

Para ilustrar a ideia, é apresentado como exemplo o trabalho realizado em um reservatório de água doce da Amazônia Legal (SILVA et al., 2013), que utilizou métodos de Mineração de Dados para extrair conhecimentos previamente desconhecidos dos pesquisadores. Os detalhes sobre a coleta dos dados e o modelo foram incluídos no Apêndice A.

Para este trabalho utilizou-se o algoritmo *Apriori* (AGRAWAL; SRIKANT et al.,

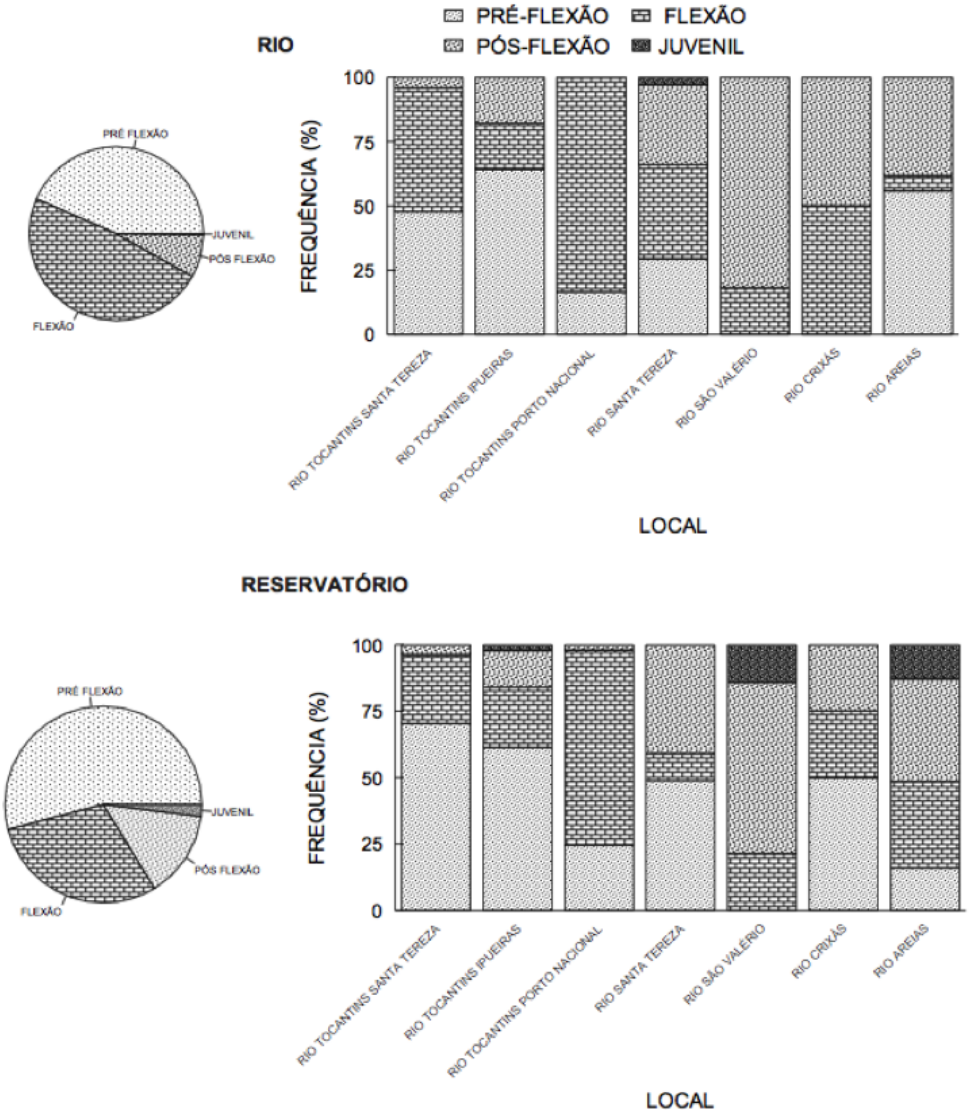

Figura 3 – Freqüência (%) das fases larvais nas Fases Rio e Reservatório (pizza) e por local (FREITAS, 2010).

1994) de regras de associação, em busca por padrões sobre uma base de dados ecológica de ictioplâncton. A concepção deste processo possibilitou extrair conhecimento do domínio da base de dados que não puderam ser descobertos somente com a utilização de métodos estatísticos (AGOSTINHO et al., 2007).

O primeiro passo foi realizar o pré-processamento dos dados para que eles pudessem ser tomados como entrada na ferramenta *Waikato Environment for Knowledge Analysis* (Weka) (HALL et al., 2009). Mais detalhes sobre essa e outras ferramentas de Mineração de Dados podem ser encontradas no Apêndice B.

O objetivo foi encontrar relacionamentos ocultos entre fatores bióticos e abióticos,

utilizando-se regras de associação geradas para este propósito. As regras desejadas seriam as que possuíssem melhor cobertura e por cobertura entendesse que ela é tida como o número de instâncias para as quais a regra prevê corretamente, chamada de suporte. Sua precisão, chamada de confiança, é o número de instâncias que a regra prevê corretamente, expressa como uma percentagem de todas as instâncias a que se aplica (WITTEN; FRANK; HALL, 2011).

Os valores utilizados para os parâmetros de suporte foram mantidos em 0.1 para suporte mínimo e 1.0 para suporte máximo, fazendo com que as regras fossem geradas livremente. Os valores dos parâmetros confiança e interesse foram alterados de 0.8 a 1.0 e de 1.5 a 3.5, respectivamente. No total, 979 regras de associação foram geradas, conforme a Figura 4.

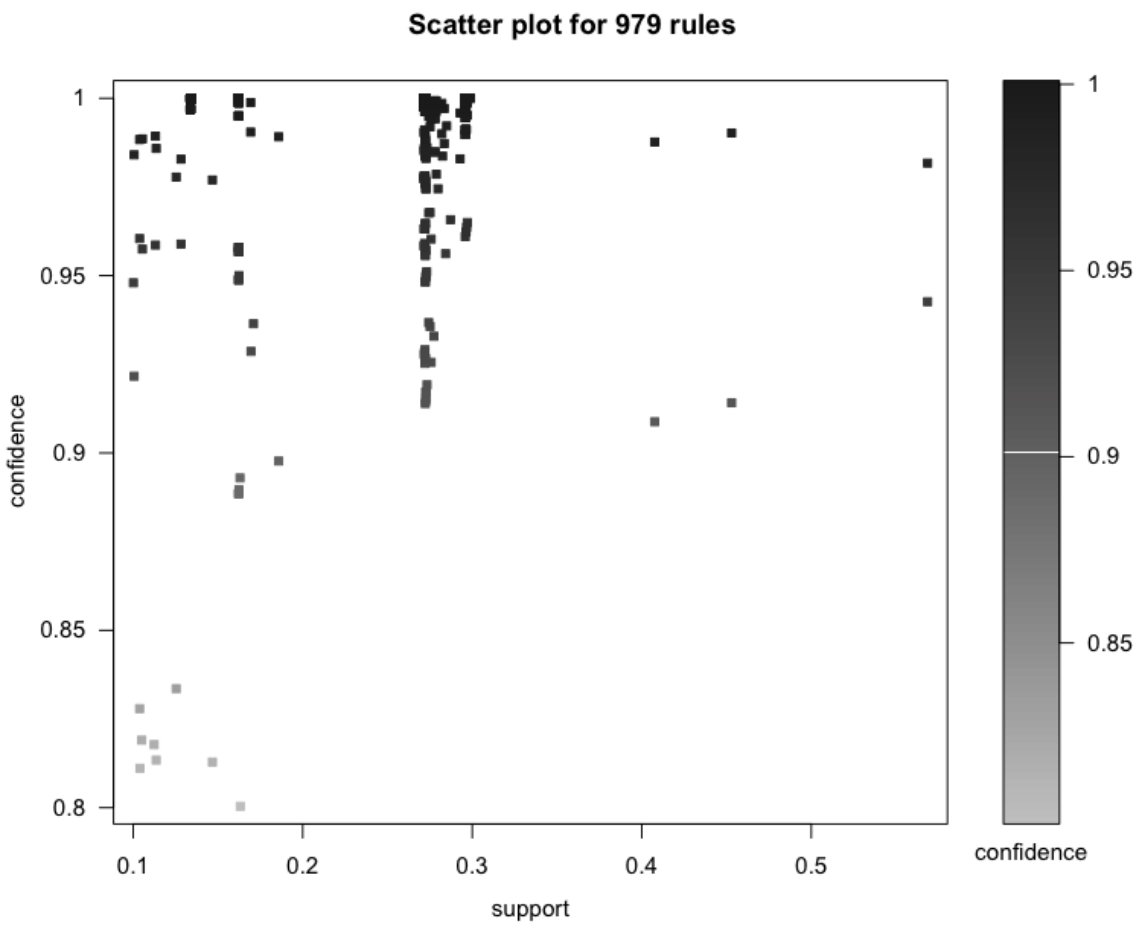

Figura 4 – Relação entre a confiança e o suporte para as 979 regras de associação geradas e seus limites. (SILVA et al., 2013).

Com o auxílio do gráfico gerado pode-se identificar as melhores regras com base na confiança e no suporte das regras. É exibido a Tabela 1, abaixo, que abrange o atributo estágio para as regras 17 e 22, com um valor de confiança bastante alto, de 0.99 e 0.96.

Estas duas regras de associação foram validadas por especialistas e o resultado considerado como conhecimento relevante sobre o assunto do trabalho em questão. Estes padrões foram extraídos com conhecimento mínimo sobre o domínio e apresentam o

Tabela 1 – Regras que envolvem o atributo estágio.

17. transparencia=6 estagio=pre 524 ==> profundidade=7 518 conf:(0.99)
22. profundidade=7 estagio=pre 541 ==> transparencia=6 518 conf:(0.96)

potencial da descoberta de conhecimento e análise exploratória de bases de dados através da mineração de dados.

Este exemplo mostra que dentre os métodos existentes de Mineração de Dados não há uma receita única para todos os tipos de dados, mas, com ferramentas como o Weka pode-se explorar os conjuntos dados facilmente, convertê-los e testar cada método individualmente.

Trabalhar com classificadores, por exemplo, pode ser muito interessante principalmente em casos onde é desejável realizar previsões sobre novos registros. As máquinas de aprendizagem também mostram desempenho promissor sobre todos os modelos construídos. Elas são simples de aplicar e servem como uma abordagem alternativa para a análise de regressão logística ou multinomial (KRUPPA et al., 2014). Em outros casos, como nos exemplos apresentados, a extração de conhecimento tem melhores ganhos se considerados padrões sobre os dados.

4 Conclusão

O estatístico Friedman já defendia a ideia de que os estatísticos utilizassem outras metodologias da computação (como a Mineração de Dados) ao invés de simplesmente trabalhar maneiras mais convenientes de aplicarem as ferramentas estatísticas já existentes (FRIEDMAN, 1998). Friedman escreveu este artigo mais de 10 anos atrás e todas suas observações sobre a importância e o aumento da prevalência de dados e poder de computação são ainda mais verdadeiros hoje do que naquela época.

Os ecologistas vêm trabalhando com a mesma forma de modelar seus sistemas há algum tempo, os mesmos métodos são aplicados repetidas vezes ou apenas com técnicas simples sem vislumbrar o potencial escondido nos dados.

Mineração de Dados se apresenta como um complemento para os estudos em sistemas ecológicos como ferramenta para análise exploratória, e não se trata de utilizar métodos de mineração de dados ou métodos de análise confirmatória exclusivamente, mas sim da utilização de ambas as metodologias para construir modelos melhores. Descobertas realizadas aplicando os métodos descritos neste trabalho podem gerar novas hipóteses e lançar um novo olhar sobre os dados.

Espera-se que novos trabalhos em sistemas ecológicos utilizem ainda mais mineração de dados e que o façam com alguns cuidados como a validação cruzada para verificação do desempenho, uso de conjuntos de testes e que no pior caso possa servir como referência para análises estatísticas, para confrontar resultados, validando-os.

Referências

AGOSTINHO, A. A. et al. Fish ladder of lajeado dam: migrations on one-way routes? *Neotropical Ichthyology*, SciELO Brasil, v. 5, n. 2, p. 121–130, 2007. Citado 2 vezes nas páginas 6 e 11.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499. Citado na página 6.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, Springer, v. 36, n. 1-2, p. 105–139, 1999. Citado na página 2.

BHATTACHARYYA, G. K. Statistical concepts and methods, gouri k. bhattacharyya, richard a. johnson. Wiley. New York. US, 1977. Citado na página 5.

BOLKER, B. M. *Ecological models and data in R*. [S.l.]: Princeton University Press, 2008. Citado na página 5.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 2.

DENNIS, B. Statistics and the scientific method in ecology (with commentary). *The nature of scientific evidence: statistical, philosophical, and empirical considerations. University of Chicago Press, Chicago, Illinois, USA*, p. 327–378, 2004. Citado na página 1.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado 2 vezes nas páginas 2 e 3.

FREITAS, I. S. *Efeito do represamento na abundância e distribuição de peixes e suas larvas numa área fluvial e a montante da Usina Hidrelétrica Luís Eduardo Magalhães, Rio Tocantins*. Dissertação (Mestrado) — Ciências do Ambiente, Universidade Federal do Tocantins, 2010. Citado 3 vezes nas páginas 5, 6 e 10.

FRIEDMAN, J. H. Data mining and statistics: What's the connection? *Computing Science and Statistics*, PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, v. 29, n. 1, p. 3–9, 1998. Citado na página 8.

HAHSLER, M. et al. The arules r-package ecosystem: analyzing interesting patterns from large transaction data sets. *The Journal of Machine Learning Research*, JMLR. org, v. 12, p. 2021–2025, 2011. Citado na página 12.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado 2 vezes nas páginas 6 e 12.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan kaufmann, 2006. Citado na página 3.

HASTIE, T. et al. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, Springer, v. 27, n. 2, p. 83–85, 2005. Citado na página 2.

HOCHACHKA, W. M. et al. Data-mining discovery of pattern and process in ecological systems. *The Journal of Wildlife Management*, Wiley Online Library, v. 71, n. 7, p. 2427–2437, 2007. Citado 2 vezes nas páginas 1 e 5.

HONG, T.-P.; KUO, C.-S.; CHI, S.-C. Mining association rules from quantitative data. *Intelligent data analysis*, IOS Press, v. 3, n. 5, p. 363–376, 1999. Citado na página 3.

KRUPPA, J. et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal*, Wiley Online Library, 2014. Citado na página 8.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery and data mining. In: *Data Mining and Knowledge Discovery Handbook*. [S.l.]: Springer, 2010. p. 1–15. Citado na página 4.

MAKLER-PICK, V. et al. Sensitivity analysis for complex ecological models–a new approach. *Environmental Modelling & Software*, Elsevier, v. 26, n. 2, p. 124–134, 2011. Citado na página 5.

SILVA, M. d. A. et al. Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In: *Advanced Data Mining and Applications*. [S.l.]: Springer, 2013. p. 384–395. Citado 3 vezes nas páginas 5, 7 e 11.

SYMONDS, M.; MOUSSALLI, A. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology*, v. 65, n. 1, p. 13–21, 2011. ISSN 0340-5443. Citado na página 5.

TUKEY, J.; BRILLINGER, D. *The collected works of John W. Tukey, philosophy and principles of data analysis 1965–1986. Volume IV*. [S.l.]: Belmont, CA: Wadsworth Advanced Books & Software, 1986. Citado na página 4.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Amsterdam: Morgan Kaufmann, 2011. ISBN 978-0-12-374856-0. Citado 2 vezes nas páginas 2 e 7.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, v. 14, n. 1, p. 1–37, 2008. Citado na página 4.

APÊNDICE A – Coleta de Dados

Para apresentar dados e resultados de pesquisas de ecologia que foram considerados neste estudo, dois são destacados como principais. Para o levantamento que utiliza análise estatística o trabalho "Efeito do represamento na abundância e distribuição de peixes e suas larvas numa área fluvial e a montante da Usina Hidrelétrica Luís Eduardo Magalhães, Rio Tocantins" (FREITAS, 2010), e para a pesquisa sobre mineração de dados o trabalho

"*Exploring an Ichthyoplankton Database from a Freshwater Reservoir in Legal Amazon*" (SILVA et al., 2013).

Os dados utilizados em ambos os trabalhos foram são fruto do trabalho "Fish ladder of Lajeado Dam: migrations on oneway routes?" (AGOSTINHO et al., 2007) produzido em razão do monitoramento da fauna de peixes na área de influência da Usina Hidrelétrica do Lajeado, na bacia do Rio Tocantins e que gerou uma das primeiras bases sistematizadas de informação do grupo de pesquisa na Amazônia Legal, com dados coletados antes, durante e após o enchimento do reservatório.

A coleta de dados de larvas de peixes foi realizada mensalmente no período de abril de 2010 a março de 2012, em doze pontos localizados no reservatório, apresentados na Figura 5.

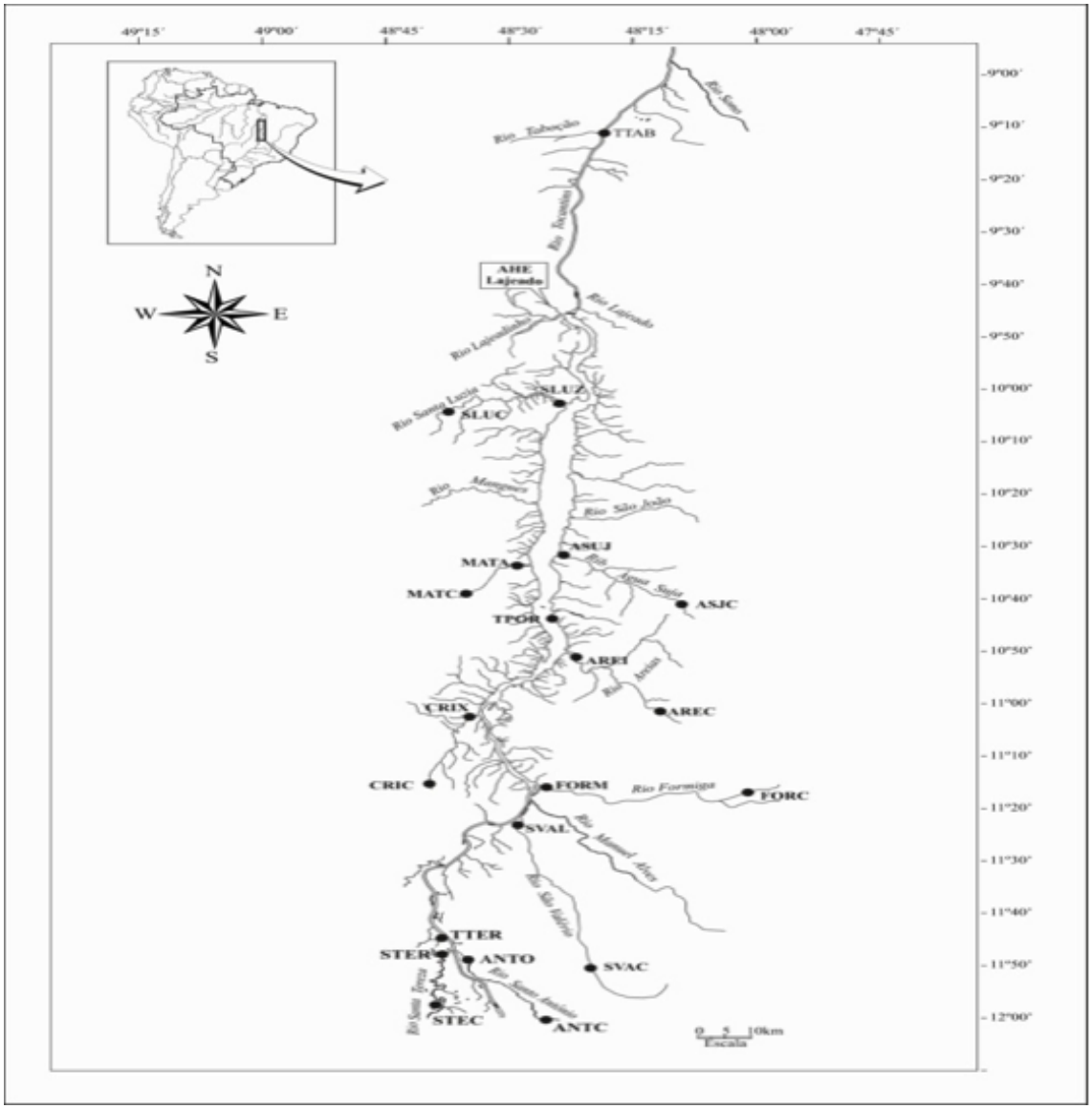

Figura 5 – Localização dos pontos de amostragem no reservatório do Lajeado e sua área de influência. . (AGOSTINHO et al., 2007).

APÊNDICE B – Ferramentas utilizadas

Algumas ferramentas fizeram parte desta pesquisa para que fosse possível realizar a mineração de dados e a apresentação dos resultados graficamente, conforme especificação neste apêndice.

Dentre as ferramentas para a mineração de dados disponíveis atualmente, a utilizada foi a *Waikato Environment for Knowledge Analysis* (Weka) (HALL et al., 2009) e para a apresentação gráfica o programa R, ambas distribuídas gratuitamente e com código fonte aberto.

O Weka é um conjunto de algoritmos para aprendizado de máquina e tarefas de mineração de dados que oferece a funcionalidade de executar estes algoritmos diretamente sobre um conjunto de dados, através da própria ferramenta, ou ainda ser executada através de um código Java. Além disso, contém ferramentas para pré-processamento de dados e oferece, intuitivamente, formas de minerar dados com classificação, regressão, clusterização e regras de associação. Desenvolvido em Java na Universidade de Waikato, Nova Zelândia, esta ferramenta pode ser encontrada em http://www.cs.waikato.ac.nz/ml/index.html.

O ambiente R, importante ferramenta que, conectada a um ambiente de dados, permite a realização de cálculos estatísticos e gráficos. É amplamente expansível através de seus pacotes, que aumentam a abrangência de cálculos disponíveis, além das funcionalidades de avaliação de resultados e interpretação de procedimentos estatísticos.

No trabalho de mineração de dados apresentado nas seções anteriores, a ferramenta R foi utilizada para a geração da apresentação gráfica dos resultados, com a extensão do pacote arules que, dentre outras funcionalidades, fornece infraestrutura para representação de padrões e regras de associação (HAHSLER et al., 2011).

Criada por Ross Ihaka e Robert Gentleman na universidade de Auckland, Nova Zelândia, a ferramenta pode ser encontrada em http://cran.r-project.org/.

7 Conclusões

Existem muitas técnicas para o pré-processamento de dados como limpeza, integração, redução e transformação dos dados e essas técnicas não são mutuamente exclusivas, podem ser aplicadas em uma mesma base de dados como por exemplo uma limpeza que envolva transformações para corrigir erros, como a transformação de campos númericos em atributos nominais através da discretização, tudo depende do problema e dos dados.

Embora tenham sido desenvolvidos vários métodos de pré-processamento de dados, ela continua sendo uma área ativa de pesquisa, devido à enorme quantidade de dados inconsistentes ou "sujos" e da complexidade dos problemas.

Neste trabalho, procurou-se demonstrar que o pré-processamento vai além das técnicas habituais, o pesquisador deve análisar os dados e manter o foco no problema, analisando entradas, custo computacional e resultados. Muitas pesquisas possuem o perfil para utilizar a Mineração de Dados como forma de explorar suas bases de dados antes mesmo de formularem hipóteses, permitindo que mesmo não especialistas no problema estudado possam ter condições para explorar e extrair algum conhecimento de bases de dados.

O processo para modelagem algorítmica descrito nesse trabalho pode ser aplicado para outros conjuntos de dados oriundos de pesquisas em que há algum tipo de coleta de dados e com isso a formação dos conjuntos como foi demonstrado pelos artigos publicados presentes neste trabalho. Em particular, para pesquisas interdisciplinares, sugere-se como trabalho futuro consolidar o processo apresentado na seção 3.3 com o IRPM.

Mais especificamente, no artigo Aplicação do algoritmo "Apriori para uma base de dados de ictioplâncton em um reservatório de água doce da Amazônia Legal", Capítulo 4, apresentou-se a aplicação de técnicas de Mineração de Dados para realização de análise exploratória de um conjunto de dados de ictioplâncton. Essa aplicação somente foi possível após a realização do pré-processamento dos dados, uma vez que os dados estavam divididos em duas bases de dados e apresentavam uma série de problemas como dados ausentes, inconsistências, redundâncias, valores errados e ruídos, que comprometiam a execução dos algoritmos de Mineração de Dados e consequentemente a qualidade dos resultados. Após a realização da fase do pré-processamento, o algoritmo Apriori foi aplicado para geração das regras de associação, o modelo algorítmico foi criado e avaliado até que se atingiu uma meta estabelecida para a confiança das regras geradas estipulada em 0.9 (em uma escala que vai de 0 até 1.0). Quando a meta não era atingida os dados eram novamente pré-processados e novamente o modelo era gerado, processo que se repetiu até que a confiança das regras atingisse o valor da meta. Concluiu-se que a Mineração

de Dados pode ser utilizada como suporte a análise exploratória dos dados, desde que os dados passem pela fase do pré-processamento. O estudo proporcionou a descoberta de conhecimento singular para o grupo de pesquisa em ictiofauna da Amazônia Legal.

Já no artigo "Explorando dados sociais para entender o trabalho infantil", Capítulo 5, apresentou-se novamente uma aplicação do processo executado no artigo anterior, dessa vez sobre uma base de dados do Programa de Assistência Social Cadastro Único (CadÚnico) com o objetivo de encontrar indicadores para a prática do trabalho infantil. Para alcançar o objetivo, aplicou-se um método de classificação utilizando árvores de decisão o que em um primeiro momento falhou, já que os próprios dados, oriundos de fontes diferentes apresentavam muitos problemas como dados ausentes, redundantes, valores errados, ruídos impedindo a criação do modelo algorítmico. Essa limitação foi superada aplicando as etapas de integração, limpeza, redução e transformação dos dados da fase do pré-processamento. A árvore de decisão gerada aponta o atributo "Frequência Escolar" como sendo um indicador mais confiável do que o atributo "Renda Familiar", isso levanta novos questionamentos em relação as políticas empregadas pela Assistência Social do Governo Federal, constituindo dessa forma um conhecimento válido perante especialistas em assistência social do Governo do Tocantins que análisaram os resultados. Dessa maneira, conclui-se que o processo utilizado no primeiro artigo também foi exequível quando aplicado sobre dados socioeconômicos de famílias brasileiras de baixa renda reunidos na base de dados do CadÚnico.

No terceiro artigo "Mineração de Dados como ferramenta de análise exploratória em Sistemas Ecológicos", Capítulo 6, apresentou-se a Mineração de Dados como um recurso importante para realização da análise exploratória em sistemas ecológicos apresentando a distinção entre ela e a análise confirmatória. Através de um comparativo entre dois trabalhos realizados sobre uma mesma base de dados o artigo expõe os resultados obtidos através de métodos estatísticos tradicionais e realizada por especialistas área, e um segundo trabalho realizado utilizando técnicas de Mineração de Dados e pré-processamento, sem o conhecimento de domínio apropriado da área. Ao final, conclui-se que as técnicas citadas são complementares para a análise exploratória de sistemas ecológicos e que a Mineração de Dados tem um grande potencial na área de Sistemas Ecológicos para construção de modelos cada vez mais eficientes.

Por fim, como trabalho futuro pretende-se realizar um comparativo que busque colocar as duas abordagens, modelagem de dados e modelagem algorítmica, lado a lado e observar os pontos em comum, suas principais diferenças, além de chamar atenção para algumas dificuldades associadas no emprego do modelo de dados.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Database mining: A performance perspective. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 5, n. 6, p. 914–925, 1993. Citado na página 25.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499. Citado na página 25.

BECKER, R. A. A brief history of s. *cahier de recherche, AT&T Bell Labora*, 1994. Citado na página 40.

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J.; OLSHEN, R. A. *Classification and regression trees*. [S.l.]: CRC press, 1984. Citado na página 33.

BREIMAN, L. et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. Citado 2 vezes nas páginas 15 e 16.

CAETANO, M. G. *Aplicando o MVC EA-IRPM ao SIGD como estratégia de evolução de sistemas Web*. Dissertação (Mestrado) — Universidade Federal do Tocantins - UFT, 2012. Citado na página 36.

CAMILO, C. O.; SILVA, J. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 1–29, 2009. Citado na página 26.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. Advances in knowledge discovery and data mining. the MIT Press, 1996. Citado 4 vezes nas páginas 15, 20, 21 e 34.

FERNANDEZ, G. *Data mining using SAS applications*. [S.l.]: CRC press, 2002. Citado na página 17.

GENG, L.; HAMILTON, H. J. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, ACM, v. 38, n. 3, p. 9, 2006. Citado na página 25.

GHISELLI, E. E. *Theory of psychological measurement*. [S.l.]: McGraw-Hill New York, 1964. Citado na página 24.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado 2 vezes nas páginas 24 e 40.

HALL, M. A. *Correlation-based feature selection for machine learning*. Tese (Doutorado) — The University of Waikato, 1999. Citado na página 24.

HALL, M. A. Feature Selection for Discrete and Numeric Class Machine Learning. In: *Machine Learning. Proc. Seventeenth International conference on Machine Learning*. [s.n.], 2000. p. 359–366. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.6025>. Citado na página 24.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. [S.l.]: Morgan kaufmann, 2006. Citado 2 vezes nas páginas 17 e 23.

HONG, T.-P.; KUO, C.-S.; CHI, S.-C. Mining association rules from quantitative data. *Intelligent data analysis*, IOS Press, v. 3, n. 5, p. 363–376, 1999. Citado 2 vezes nas páginas 15 e 21.

HUNT, E. B.; MARIN, J.; STONE, P. J. Experiments in induction. Academic Press, 1966. Citado na página 30.

KANTARDZIC, M. *Data mining: concepts, models, methods, and algorithms*. [S.l.]: John Wiley & Sons, 2011. Citado na página 17.

LAKSHMI, N.; RAO, R. S.; REDDY, S. S. An overview of preprocessing on web log data for web usage analysis. *International Journal of Computer Applications. India*, v. 2, n. 4, p. 274–279, 2013. Citado na página 35.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado na página 21.

LETOUZE, P. Interdisciplinary research project management. *International Proceedings of Economics Development & Research*, v. 14, 2011. Citado 4 vezes nas páginas 10, 36, 37 e 38.

LETOUZE, P.; CAETANO, M. G.; ISHIHARA, J. Y.; PRATA, D.; BRITO, G. Applying mvc to evolutionary acquisition irpm. *International Proceedings of Computer Science & Information Technology*, v. 45, 2012. Citado na página 36.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery and data mining. In: *Data Mining and Knowledge Discovery Handbook*. [S.l.]: Springer, 2010. p. 1–15. Citado 2 vezes nas páginas 10 e 22.

MÁRQUEZ-VERA, C.; CANO, A.; ROMERO, C.; VENTURA, S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, Springer, v. 38, n. 3, p. 315–330, 2013. Citado na página 35.

MOORE, J. H.; RITCHIE, M. D. The central role of biological data mining in connecting diverse disciplines. *BioData mining*, BioMed Central, v. 6, p. 14, 2013. Citado na página 16.

NASA. *The Open Data Project of NASA*. 2014. Disponível em: <http://data.nasa.gov>. Citado na página 16.

ORACLE. *The Java Programming Language and the Java Platform*. 2013. Disponível em: <http://www.oracle.com/technetwork/topics/newtojava/downloads/index.html>. Citado na página 40.

PIERRAKOS, O.; ZILBERBERG, A.; ANDERSON, R. Understanding undergraduate research experiences through the lens of problem-based learning: Implications for curriculum translation. *Interdisciplinary Journal of Problem-based Learning*, v. 4, n. 2, 2010. Citado na página 36.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Morgan kaufmann, 1993. Citado na página 27.

QUINLAN, J. R. et al. *Discovering rules by induction from large collections of examples*. [S.l.]: Expert systems in the micro electronic age. Edinburgh University Press, 1979. Citado na página 30.

RODRIGUES, D. C.; PRATA, D. N.; SILVA, M. A. Exploring social data to understand child labor. *International Journal of Social Science and Humanity*, v. 5, n. 1, p. 29–33, 2013. Citado 2 vezes nas páginas 19 e 39.

SAVERY, J. R. Overview of problem-based learning: Definitions and distinctions. *Interdisciplinary Journal of Problem-based Learning*, v. 1, n. 1, p. 3, 2006. Citado na página 36.

SILVA, M. d. A.; TREVISAN, D. Q.; PRATA, D. N.; MARQUES, E. E.; LISBOA, M.; PRATA, M. Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In: *Advanced Data Mining and Applications*. [S.l.]: Springer, 2013. p. 384–395. Citado 3 vezes nas páginas 18, 38 e 39.

TEAM, R. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2012*. [S.l.]: ISBN 3-900051-07-0,[Available from: http://www. R-project. org/], 2013. Citado na página 40.

WEIR, N.; FAYYAD, U. M.; DJORGOVSKI, S.; RODEN, J. The skicat system for processing and analyzing digital imaging sky surveys. *Publications of the Astronomical Society of the Pacific*, JSTOR, p. 1243–1254, 1995. Citado na página 34.

WITTEN, I.; FRANK, E.; HALL, M. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780080890364. Citado 4 vezes nas páginas 10, 16, 20 e 27.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; PHILIP, S. Y. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, v. 14, n. 1, p. 1–37, 2008. Citado 2 vezes nas páginas 25 e 29.

YANG, Q.; WU, X. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, World Scientific, v. 5, n. 04, p. 597–604, 2006. Citado na página 16.

ZHU, X.; DAVIDSON, I. *Knowledge Discovery and Data Mining: Challenges and Realities*. [S.l.]: Igi Global, 2007. Citado na página 35.