MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

RAFAEL CASTANEDA RIBEIRO

UM AMBIENTE DE IMPUTAÇÃO SEQÜENCIAL PARA CENÁRIOS MULTIVARIADOS

Rio de Janeiro
Novembro de 2008

INSTITUTO MILITAR DE ENGENHARIA

RAFAEL CASTANEDA RIBEIRO

UM AMBIENTE DE IMPUTAÇÃO SEQÜENCIAL PARA CENÁRIOS MULTIVARIADOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. Ricardo Choren Noya - D.Sc.
Co-orientador: Prof. Ronaldo Ribeiro Goldschmidt - D.Sc.

Rio de Janeiro
Novembro de 2008

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

R484a Castaneda,R.

Um Ambiente de Imputação Seqüencial para Cenários Multivariados/ Rafael Castaneda Ribeiro. – Rio de Janeiro: Instituto Militar de Engenharia, Novembro de 2008.

78 p.:il.

Dissertação (mestrado) – Instituto Militar de Engenharia – Rio de Janeiro, Novembro de 2008.

1. Valores Ausentes. 2. Inteligência Artificial. 3. Mineração de dados. I. Título. II. Instituto Militar de Engenharia.

CDD 006.3

INSTITUTO MILITAR DE ENGENHARIA

RAFAEL CASTANEDA RIBEIRO

UM AMBIENTE DE IMPUTAÇÃO SEQÜENCIAL PARA CENÁRIOS MULTIVARIADOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. Ricardo Choren Noya - D.Sc.

Co-orientador: Prof. Ronaldo Ribeiro Goldschmidt - D.Sc.

Aprovada em 25 de Novembro de 2008 pela seguinte Banca Examinadora:

Prof. Ricardo Choren Noya - D.Sc. do IME - Presidente

Prof. Ronaldo Ribeiro Goldschmidt - D.Sc. do IME

Prof[a]. Maria Claudia Reis Cavalcanti - D.Sc. do IME

Prof. Jorge de Abreu Soares - D.Sc. do UERJ

Prof. Luis Alfredo V. de Carvalho - D.Sc. do COPPE/UFRJ

Rio de Janeiro
Novembro de 2008

À minha família, que sempre suportou minhas iniciativas, e me apoiou durante toda a duração do programa. À minha esposa, Luciana Rodrigues Lucie, por me compreender pacientemente durante as inúmeras vezes em que tive de me concentrar nos estudos.

AGRADECIMENTOS

Agradeço a Deus por todas as oportunidades que tive em minha vida.

Aos meus orientadores, Profs. Ricardo Choren e Ronaldo Ribeiro Goldschmidt, sem os quais nada seria possível. A sua paciência e competência foram fundamentais no desenvolvimento deste trabalho e para o meu próprio crescimento pessoal.

Aos meus colegas e mentores, Profs. Jorge de Abreu Soares e Cláudia Ferlin, por estarem sempre ao meu lado, com apoio fundamental para que eu não me desviasse , nem desistisse, do caminho que escolhi.

A todos os professores e funcionários do Departamento de Sistemas e Compuação do IME, que de alguma forma, contríbuiram para a realização deste trabalho.

Aos Professores membros da banca, por terem aceito o convite, e contribuído de forma tão positiva para a conclusão e enriquecimento deste projeto.

Rafael Castaneda Ribeiro

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ABREVIATURAS

API	-	*Application Program Interface*
ARFF	-	*Attribute-Relation File Format*
CPU	-	*Central Processing Unit*
CSV	-	*Comma Separated Values*
EM	-	*Expectation-Maximization*
JGAP	-	*Java Genetic Algorithms Package*
JOONE	-	*Java Object-Oriented Neural Networks*
KDD	-	*Knowledge Discovery in Databases*
KNN	-	*k-Nearest Neighbors*
kNN-IC	-	*k-Nearest Neighbors for Incomplete Cases*
MAR	-	*Missing at Random*
Mb	-	*Megabytes*
MCAR	-	*Missing Completely at Random*
MICE	-	*Multiple Imputation by Chained Equations*
NDSL	-	*National Science Digital Library*
NMAR	-	*Not Missing at Random*
PSO	-	*Particle Swarm Optimization*
RAD	-	*Relative Absolute Deviation*
SAS	-	*Statistical Analysis System*
SGBD	-	*Sistema Gerenciador de Banco de Dados*
SRMI	-	*Sequential Regression Multiple Imputation*
TI	-	*Tecnologia da Informação*
WFMS	-	*Workflow Management System*
XML	-	*Extensible Markup Language*

ABSTRACT

Data analisys tasks usually face the problem of missing values, especially when they occur in a multivariate distribution. Sequential imputation is a common method for the completion of data, as it imputes each attribute with missing values, one at a time. However, researchers often diverge on the pratics on sequential imputation, such as reusing imputed values between the imputation of different attributes. Such divergence comes, among other factors, from the lack of established tools and methodologies that are able to provide the basis for experimentation and assessment on sequential imputation. This dissertation provides a theoretical and practical approach on sequential imputation, with the proposal of a methodology for imputation execution and assessment, and a implementation of the methodology as a workflow-based imputation environment. This work shows that the environment is capable of automating several experiments, which would be otherwise manually driven, one at time. A study case is conducted on the environment, in order to explore open questions regarding the reuse of values in sequential imputation, and to demonstrate its main features and capabilities. The study shows that the reuse of values was able to improve the overall accuracy of imputation procedures, in most of the performed experiments.

RESUMO

Uma dificuldade comum aos processos de descoberta de conhecimento em bases de dados é a existência de valores ausentes, especialmente quando estes ocorrem de maneira multivariada. O procedimento de imputação seqüencial é uma abordagem comum para a complementação dos dados, imputando os valores ausentes em cada atributo, um por vez. Porém, pesquisadores comumente divergem sobre as práticas de imputação seqüencial, como a reutilização de valores imputados entre a imputação de diferentes atributos. Tal divergência é fruto, dentre outros fatores, da falta do estabelecimento de metodologias e ferramentas capazes de fornecer as bases para experimentação e avaliação dos métodos de imputação seqüencial. Esta dissertação oferece uma abordagem teórica e prática em imputação seqüencial, com a proposta de uma metodologia para execução e avaliação de métodos de imputação seqüencial, e a implementação da metodologia na forma de um ambiente de imputação baseado nos principais conceitos de workflows. Este trabalho mostra que o ambiente é capaz de automatizar diversos experimentos, que seriam normalmente conduzidos manualmente, um-a-um. Um estudo de caso é conduzido no ambiente, para explorar questões em aberto relativas ao reuso de valores em processos de imputação seqüencial, e para demonstrar as principais funcionalidades do ambiente. Este estudo de caso mostra que o reuso de valores foi capaz de aprimorar a precisão dos procedimentos de imputação, na maior parte dos experimentos realizados.

1 INTRODUÇÃO

Diversas bases de dados, tanto na indústria quanto em centros de pesquisa, apresentam a ocorrência de valores ausentes (FARHANGFAR, 2007). Um exemplo é a base de dados da National Science Digital Library (NSDL) (NSDL.ORG, 2008) é uma coleção de recursos científicos, como artigos ou vídeos educacionais. Cada recurso é associado a até noventa campos de meta-informação, como o nome do autor, ou a área de pesquisa relacionada com o recurso, que auxiliam no processo de busca e recuperação dos recursos desejados. Nesta base de dados, 23% das entradas são valores ausentes do campo "assunto" (YI, 2007).

Valores ausentes são um obstáculo aos processos de análise de dados (SCHAFER, 1997). Tomando como exemplo a base de dados NSDL, onde o campo "assunto" é um dos mais pesquisados, a ausência de valores prejudica sua utilização potencial. Há varias causas para este problema, tais como falha humana em processos manuais de entrada de dados, erros de equipamentos e sensores, preenchimento incompleto de formulários, falhas e erros em sistemas gerenciadores de bancos de dados, entre outras (MONARD, 2003).

De um modo geral, existem duas abordagens para lidar com o problema da análise de dados em bases que apresentem valores ausentes. O analista pode remover os registros e atributos com valores ausentes do processo de análise, ou previamente imputá-los com novos valores (FARHANGFAR, 2007). A primeira opção é de mais fácil implementação, porém diversos fragmentos de informação são descartados junto com os registros, o que pode gerar resultados de análise incompletos ou inverossímeis (YUAN, 2008).

Imputação é o nome dado a um procedimento automático ou semi-automático, capaz de preencher os valores ausentes encontrados nas bases de dados (SCHÖNER, 2004) (GOLDSCHMIDT, 2005). Se apresenta como uma abordagem alternativa, a ser empregada quando o analista não deseja descartar os fragmentos de informação contidos em registros com valores ausentes (SOARES, 2007). Métodos de imputação são tarefas de complexa aplicação prática, dada a grande quantidade de técnicas disponíveis para o cálculo de valores de substituição, que variam desde operações simples como a média ou moda a modelos estatísticos e técnicas de inteligência de artificial, ou combinações de ambos (FARHANGFAR, 2007) (LAKSHMINARAYAN, 1999).

1.1 IMPUTAÇÃO UNIVARIADA E MULTIVARIADA

A ausência de valores em uma base pode ocorrer de forma univariada ou multivariada. O caso univariado se dá quando os valores ausentes estão dispostos em apenas um atributo. Neste cenário, os casos incompletos (que apresentam valores ausentes) são separados para imputação, e os casos completos são utilizados como fonte de treinamento ou consulta para as técnicas de predição de valores, como ilustrado na figura 1.1:

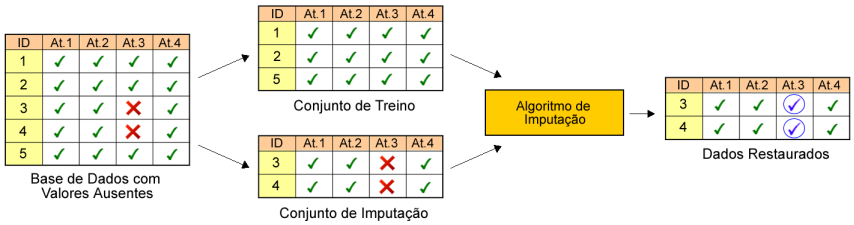


FIG. 1.1: Processo de imputação univariado

No entanto, com o rápido aumento da complexidade e do tamanho das bases de dados nos últimos anos, têm-se observado uma intensificação na pesquisa e desenvolvimento de mecanismos para imputação multivariada, quando os valores ausentes estão dispostos em dois ou mais atributos (SCHAFER, 1997). Estes casos apresentam uma série de dificuldades com as quais os mecanismos de imputação univariada são incapazes de lidar (VANBUUREN, 2006):

a) Os casos utilizados na predição de valores ausentes podem apresentar valores ausentes em outros atributos;

b) É possível encontrar casos em que um mesmo registro possui valores ausentes em dois ou mais atributos;

c) O algoritmo de imputação pode sugerir valores inconsistentes entre dois ou mais atributos, como o exemplo dos "homens grávidos" (VANBUUREN, 2006);

d) Os atributos a serem imputados possuem diferentes ordens de grandeza, o que aumenta a complexidade do problema de imputação.

Uma das principais técnicas de imputação multivariada desenvolvidas para estes cenários é a imputação seqüencial (COMISSION, 2000). A Imputação seqüencial parte de uma base de dados com valores ausentes, e realiza um processamento iterativo sobre os elementos, imputando-os com mecanismos univariados até que a base tenha todos os valores

preenchidos. A principal característica da imputação seqüencial é imputar seqüencialmente os valores ausentes, reduzindo um problema multivariado, em diversos problemas univariados. (OUDSHOORN, 1999; GELMAN, 2006).

Existem duas técnicas para a imputação seqüencial: imputação atributo-a-atributo e imputação caso-a-caso. Na imputação atributo-a-atributo (LEPKOWSKI, 2001; OUDSHOORN, 1999), os valores ausentes são processados em apenas um atributo de cada vez, mesmo em registros que apresentem dois ou mais valores ausentes. Já na imputação caso-a-caso (KIM, 2004; VERBOVEN, 2007), os valores ausentes são imputados em um registro de cada vez, independente da quantidade de atributos de valores ausentes. Esta dissertação trata somente da imputação seqüencial atributo-a-atributo, e a fim de simplificar o texto, todas as menções futuras à imputação seqüencial devem ser entendidas como atributo-a-atributo.

1.2 PROBLEMA

Quando se trabalha com a imputação seqüencial, pode-se optar por fazer reuso de valores calculados durante o processo de imputação. A idéia é, ao término da imputação de cada atributo da base, reaproveitar os valores imputados para a construção do(s) conjunto(s) de treino para a imputação do(s) próximo(s) atributo(s) (GELMAN, 2006), de maneira a melhor preservar a consistência entre os valores imputados. Uma comparação entre imputação seqüencial com e sem reuso de dados é ilustrada na figura 1.2.

Existe uma série de divergências na literatura sobre as reais vantagens ou prejuízos advindos das diversas formas de se implementar mecanismos de imputação seqüencial, como por exemplo, o reuso de valores, caracterizado pela possibilidade de se preencher os itens imputados na base dados durante o processo de imputação. Entre as questões em aberto encontradas na literatura, destacam-se as seguintes:

a) Uma distribuição muito diversa dos valores ausentes pode dificultar a identificação dos casos de treino (GELMAN, 2007).

b) A imputação seqüencial sem reuso pode levar a construção de casos inconsistentes (VANBUUREN, 2006).

c) A reutilização de valores imputados pode promover um fortalecimento artifical da correlação entre os atributos (SCHAFER, 1998).

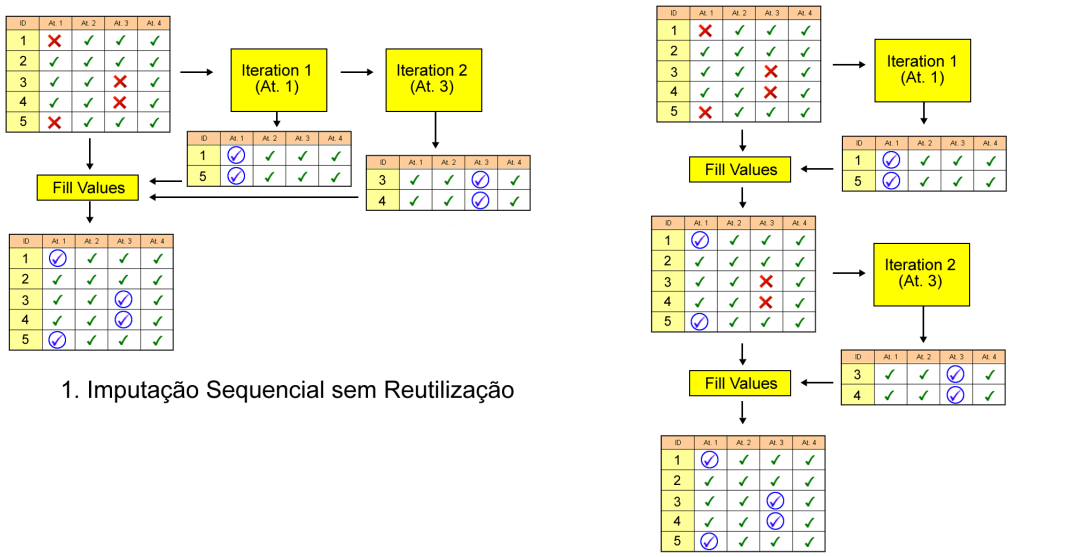


FIG. 1.2: Processo de imputação seqüencial com e sem realimentação

d) A ordem da sequência do reuso pode se tornar um fator de grande influência na qualidade final da imputação (FERLIN, 2008).

O esclarecimento destas e outras questões conflitantes é um problema relevante e significativo dentro do contexto da imputação seqüencial. Um dos maiores obstáculos deste estudo reside em como aplicar o reuso de valores, e ao mesmo tempo avaliar adequadamente o seu impacto, isolando ou amenizando outros fatores do processo de imputação. Esta dificuldade é reflexo, entre outros fatores, da carência de metodologias e ferramentas capazes de controlar o processo de imputação a fim de atender requisitos relacionados a experimentação e avaliação entre diferentes abordagens, como por exemplo, com ou sem reutilização de valores.

1.3 OBJETIVOS

Os objetivos deste trabalho são:

i Definir e implementar uma metodologia para execução e avaliação de técnicas de imputação seqüencial;

ii Utilizar a metodologia para analisar, dentro do contexto de um conjunto de bases, diferentes técnicas de imputação seqüencial, a fim de aprofundar o entendimento

das questões identificadas na literatura atual. Em especial, as relacionadas com o reuso de valores, e a escolha da ordem de imputação.

A definição de uma metodologia auxilia na comparação entre diferentes trabalhos e técnicas. Se adotada de maneira padronizada, pode proporcionar comparações justas sobre vários algoritmos de imputação, eliminando ou suavizando incertezas relativas aos métodos de experimentação, tais como a qualidade de implementação de rotinas matemáticas auxiliares (por exemplo, componentes pré-prontos do Matlab versus implementações próprias em C ou Java), diferenças entre linguagens e plataformas de programação empregadas, entre outros. Ainda, podem existir distorções entre diferentes metodologias de avaliação, capazes de beneficiar ou prejudicar o produto final da análise de resultados.

Utilizando a metodologia implementada, a análise experimental entre diferentes técnicas de imputação seqüencial procura ampliar, dentro do contexto de um conjunto de bases selecionadas, o conhecimento prático sobre algumas questões apontadas na literatura:

- Se o reuso de valores é capaz de beneficiar o processo de imputação, aprimorando os valores aferidos para substituição;
- Se o reuso de valores realmente promove um aumento artificial entre a correlação dos atributos da base imputada;
- Se o reuso de valores consegue atenuar os conhecidos efeitos negativos de uma distribuição muito esparsa dos valores ausentes;
- Se a ordem de imputação dos atributos pode influenciar significativemente o resultado final do processo.

1.4 CONTRIBUIÇÕES ESPERADAS

As contribuições esperadas deste trabalho são:

- A definição de uma metodologia para execução e avaliação de técnicas de imputação, capaz de apoiar novos trabalhos de experimentação.
- Desenvolvimento de um ambiente que implemente a metodologia proposta, utilizando conceitos de workflows.

- Apresentação de resultados experimentais que permitam concluir, no domínio de um conjunto determinado de bases de dados, se:

  - O reuso de valores pode aprimorar a qualidade final dos dados imputados;

  - O reuso de valores promove um aumento artificial da correlação entre atributos;

  - A ordem de imputação impacta de maneira significativa o processo de imputação;

1.5 ORGANIZAÇÃO DO TEXTO

O restante desta dissertação está organizado da seguinte maneira: O capítulo 2 apresenta fundamentação sobre os conceitos básicos em imputação multivariada e workflows para KDD; o capítulo 3 detalha a solução proposta e a metodologia de experimentação; o capítulo 4 relata os resultados dos experimentos realizados; o capítulo 5 aborda os trabalhos relacionados; e o capítulo 6 relaciona as conclusões, trabalhos futuros e outras considerações finais.

2 CONCEITOS BÁSICOS

A crescente complexidade dos sistemas de TI demanda o armazenamento de massas de dados cada vez maiores. Estas bases de dados comumente apresentam a ocorrência de valores ausentes ou corrompidos de informação, por motivos que variam desde erros de programação a falhas nos mecanismos de condução e preenchimento de pesquisas mercadológicas (SOARES, 2007).

Quando confrontado com um cenário como este, a opção mais simples que o analista pode escolher para lidar com valores ausentes é removê-los (MAGNANI, 2004). Esta remoção pode se dar de duas maneiras, pela análise exclusiva dos casos completos (*complete-case analisys*), ou pela análise dos casos disponíveis (*available-case analisys*).

A análise de casos completos consiste na exclusão de todos os registros que apresentam valores ausentes. Este método é o de mais fácil implementação, porém diversos fragmentos de informação são descartados junto com os registros, o que pode gerar resultados de análise incompletos ou inverossímeis (YUAN, 2008). Já na análise de casos disponíveis, existe um relaxamento que ameniza o problema do descarte de informação. São excluídos do processo de análise apenas os casos que possuem valor(es) ausente(s) nos atributos relevantes àquela etapa da análise, não necessariamente em qualquer um.

Um exemplo acontece no cálculo de correlação, uma medida que estabelece o quanto a variação dos valores de dois atributos estão relacionadas entre si, como as notas de um aluno e as horas de estudo, ou o peso e a altura. Considerando duas variáveis A e B, têm-se a EQ. 2.1, onde σ é o desvio padrão, e *cov* é a co-variância.

$$\rho_{A,B} = \frac{\mathrm{cov}(A, B)}{\sigma_A \sigma_B} \tag{2.1}$$

Assim, pode ser observado que para se calcular a correlação entre dois atributos de uma base, não é necessário excluir todos os casos que apresentem valores ausentes, mas apenas aqueles que apresentem valores ausentes em um dos dois atributos cuja correlação está sendo avaliada (COHEN, 2000).

Este é um método mais complicado que a remoção simples, uma vez que nem todos os procedimentos de análise são tão triviais quanto o cálculo de correlação. Em geral, a

análise de casos disponíveis é menos empregada que a remoção completa, uma vez que o trabalho extra de implementação e execução pode ser melhor empregado em técnicas mais sofisticadas (MAGNANI, 2004). Se o analista não deseja perder dados no tratamento de valores ausentes, ele pode optar por substituir os valores ausentes por valores estimados.

2.1 IMPUTAÇÃO SEQÜENCIAL

Imputação é qualquer procedimento automático ou semi-automático, capaz de preencher valores ausentes encontrados em bases de dados (SCHÖNER, 2004; GOLDSCHMIDT, 2005). Métodos de imputação vêm sendo pesquisados e aplicados desde a década de 70. Inicialmente restritos ao domínio da estatística, estes métodos evoluíram, e apresentam hoje implementações baseadas em inteligência artificial, ou até mesmo construções híbridas. (FARHANGFAR, 2007; LAKSHMINARAYAN, 1999).

Os processos de imputação podem ser diferenciados pela capacidade de imputar valores ausentes que ocorrem de maneira univariada ou multivariada. A imputação é dita univariada quando os valores ausentes estão dispostos em apenas um atributo. Nestes cenários, existem diversas técnicas que são comumente aplicadas. Um estudo comparativo em Imputação Univariada de dados pode ser encontrado em Soares (SOARES, 2007).

Entretanto, com o rápido aumento da complexidade e do tamanho das bases de dados nos últimos anos, têm-se observado uma intensificação na pesquisa e desenvolvimento de mecanismos para imputação multivariada, quando os valores ausentes estão dispostos em dois ou mais atributos (SCHAFER, 1997).

Existem duas maneiras de se abordar a solução de problemas multivariados (VANBUREN, 2006): modelagem conjunta (*joint-modelling*) ou especificação condicional (*conditional especification*). A modelagem conjunta consiste em utilizar modelos estatísticos (como redes de bayes) para estimar valores ausentes em todos os atributos de uma única vez.

As técnicas de especificação condicional são, na verdade, generalizadas na literatura pelo termo imputação seqüencial (COMISSION, 2000), nas quais os valores ausentes a serem imputados são processados de maneira seqüencial sobre os atributos (LEPKOWSKI, 2001; OUDSHOORN, 1999), ou sobre os registros (KIM, 2004; VERBOVEN, 2007). A principal característica de se imputar seqüencialmente os valores ausentes é a redução de um problema multivariado, em diversos problemas univariados, que podem ser resolvidos pelas já conhecidas técnicas tradicionais de imputação univariada. (OUDSHOORN, 1999;

GELMAN, 2006).

Por preservar as características de cada atributo, e permitir a flexibilização de diferentes modelos para solução de cada problema, a imputação seqüencial é considerada a melhor escolha por diversos autores (VANBUUREN, 2006; GELMAN, 2001; HEERINGA, 2002). O processo de imputação seqüencial atributo-a-atributo, que é o foco desta dissertação, está ilustrado na figura 2.1, onde uma tabela apresenta um padrão de valores ausentes multivariado nos atributos 1 e 3. Cada atributo é resolvido de maneira independente, e ao final os valores imputados são preenchidos na base de dados.

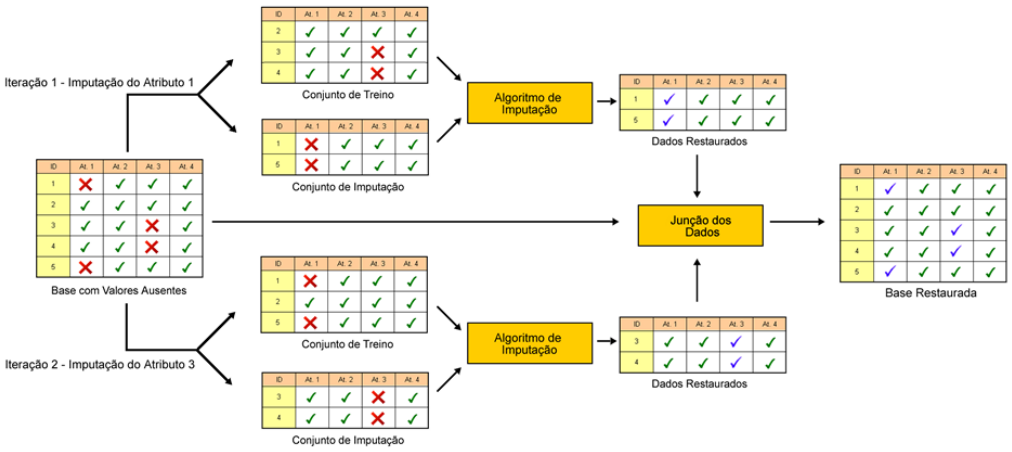


FIG. 2.1: Imputação seqüencial

Embora as técnicas de imputação seqüencial sejam aplicadas com sucesso, elas possuem suas próprias limitações e deficiências, que levam a exploração de implementações variantes (RUBIN, 2003; ROYSTON, 2005). Uma das principais deficiências da técnica tradicional de imputação seqüencial está em se fazer a divisão de um problema N-variado em N problemas univariados e independentes, de maneira que o processo perde a relação de interdependência que existe entre os atributos da base.

Por exemplo, (VANBUUREN, 2006) relata o acontecimento de casos de imputação inconsistente, como o dos "homens grávidos" . Isto aconte pois algoritmos não tratam a relação entre mais de um valor imputados para um mesmo registro. Ainda, em bases onde a ausência de valores se distribui por diversos atributos, a técnica apresenta problemas, pois os dados ausentes se espalham por todos os conjuntos de treino e influenciam na imputação dos atributos (GELMAN, 2006).

Uma das soluções é a aplicação de uma técnica denominada reuso de valores (KIM,

2004; GELMAN, 2006), onde os valores imputados para um atributo ou registro da sequência são inseridos na base de dados antes da próxima imputação tomar lugar, possivelmente colaborando nos futuros conjuntos de treino. A figura 2.2 ilustra a imputação seqüencial com reuso:

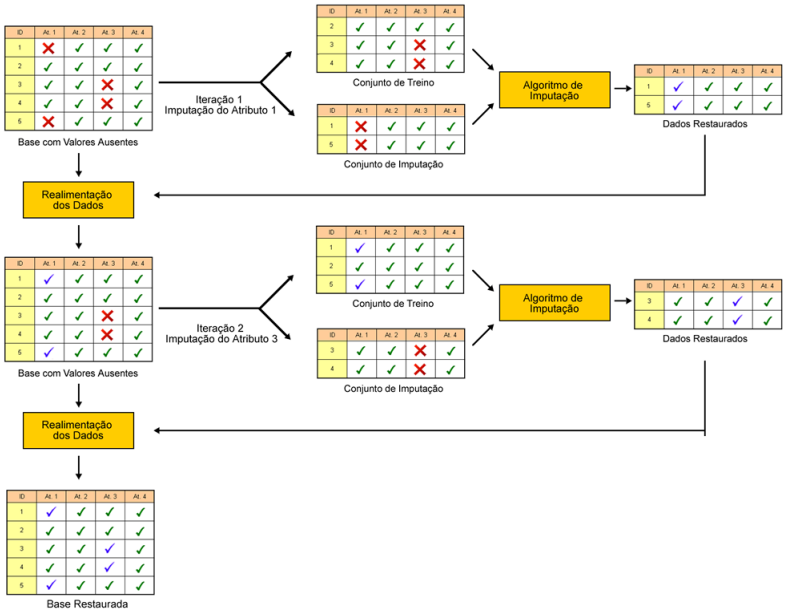


FIG. 2.2: Imputação seqüencial com Reuso

Por exemplo, se o atributo Ax possui um alto grau de correlação com o atributo Ay, e os dois apresentam uma quantidade razoável de valores ausentes, provavelmente nenhum dos dois atributos será adequadamente restaurado, ou, pelo menos, tão bem restaurados quanto poderiam ser, caso os valores descobertos para Ax fossem reaproveitados na próxima iteração, servindo de entrada para a inferência dos valores ausentes em Ay.

Ao aplicar este reuso de valores entre as iterações, é possível conseguir um progressivo enriquecimento do conjunto de treino, que fica mais completo a cada iteração, o que pode ser muito útil em casos onde há uma ampla ocorrência de valores ausentes. Ainda, a relação entre os atributos é preservada. Por exemplo, um registro que teve o valor ausente de "gênero" imputado como "masculino", não poderia, após a realimentação deste valor, ter o valor "verdadeiro" para o atributo "gravidez". Encadeando a resolução dos atributos, procura-se preservar a consistência do processo de imputação iterativa, e progressivamente diminuir a quantidade de valores ausentes encontrados nos conjuntos de treino. Outro exemplo é relatado em (LEPKOWSKI, 2001), onde se estabelece a dependência entre as variáveis "idade" e "anos de fumante" de um grupo de pessoas.

No campo da imputação multivariada de dados, trabalho de Gleason (GLEASON, 1974) foi um dos primeiros a sugerir que valores imputados para uma coluna sejam utilizados para recalcular métricas estatísticas sobre a base de dados, a fim de auxiliar em futuras iterações do processo de imputação. No entanto, o trabalho não menciona a possibilidade de reutilização direta.

Já o algoritmo *Multivariate Imputation by Chained Equations* (MICE) (OUDSHOORN, 1999) trabalha com a construção de equações encadeadas, onde a imputação de um campo é considerada uma variável cuja resolução é dada por uma fórmula. Quando os valores ausentes se apresentam de maneira multivariada, as equações para cada variável são encadeadas, de maneira que a resolução de um atributo alimenta a equação e, por conseguinte, a resolução de outro. Segundo os próprios autores, dada a natureza extremamente encadeada da resolução do problema, o algoritmo não apresenta bom desempenho quando os valores ausentes ocorrem em muitas colunas, fato que prejudica a construção e o encadeamento das equações.

Na reutilização efetiva de valores, porém, o algoritmo *Sequential Regression Multivariate Imputation* (SRMI) foi o pioneiro (LEPKOWSKI, 2001). Este algoritmo funciona construindo modelos preditivos para cada coluna da base de dados, começando da coluna com menos valores ausentes para a coluna com mais valores ausentes. A cada iteração, os valores descobertos pelo modelo preditivo anterior são aproveitados na geração do próximo modelo preditivo. O SRMI trabalha com o conceito de "restrictions and bounds", informações de conhecimento do analista de dados sobre os valores limítrofes que podem ser encontrados nas colunas, e regras auxiliares que limitam a ação dos modelos preditivos, por exemplo: "se idade < 18 então anos_fumante $= 0$".

Segundo os autores, muito embora a imputação seja possível sem a adição destas regras, o desempenho apresentado nestes casos é sempre inferior. Esta diferença deve ser encarada com cautela, pois ao mesmo tempo em que a introdução de conhecimento prévio do analista pode ajudar na imputação, esta "necessidade" torna o algoritmo semi-automático e frágil, a partir do momento em que o analista forneça informações errôneas, ou simplesmente não conheça a fundo o domínio da base de dados.

Existem diversos outros trabalhos, como por exemplo (KENNICKELL, 1997; FARIS, 2002; HEERINGA, 2002; JONSSON, 2004), que tratam a resolução de problemas de ausência multivariada de dados, mas não lidam com qualquer mecanismo de reutilização de valores durante o processo de imputação. Um resumo comparativo de alguns destes

trabalhos pode ser encontrado em (GELMAN, 2007).

2.2 ALGORITMOS DE IMPUTAÇÃO

De acordo com o glossário de termos estatísticos das Nações Unidas (COMISSION, 2000), existem seis categorias principais para a classificação dos algoritmos de imputação:

a) Imputação Determinística - O algoritmo de imputação é determinístico quando existe uma resposta única e correta a ser obtida para o valor ausente. Um exemplo simples são valores de totalização ou sumarização que se encontram ao final de colunas numéricas, e que podem ser facilmente recalculados.

b) Imputação Baseada em Modelos - A imputação em modelos é uma categoria diversa, relacionada a métodos estatísticos. Pertencem a ela implementações simples, como o preenchimento pela média ou moda dos atributos, e abordagens mais sofisticadas, como regressões lineares e o algoritmo EM (expectation-maximization).

c) Deck-Based Imputation - Neste tipo de imputação, o valor para preenchimento em um registro com dado ausente é derivado dos valores presentes em diversos outros registros "doadores" (sejam eles em si casos completos ou não). Se os casos doadores estiverem presentes na mesma base de dados que o caso ausente, a imputação é determinada *hot-deck*. Caso contrário, eles podem ser encontrados em outras bases do mesmo domínio (como históricos e cópias de segurança), quando a imputação é denominada *cold-deck*. O algoritmo dos "vizinhos mais próximos" (*Nearest Neighbors*) é uma das técnicas de busca mais empregadas para o reconhecimento de casos de doadores que guardem a maior semelhança possível com o caso ausente.

d) Imputação Híbrida - Técnicas de imputação que combinam diferentes estratégias. Como exemplos podem ser citados a Imputação Composta (SOARES, 2007) e a Imputação em Cascata (FERLIN, 2008).

e) Sistemas Especialistas - Programas de computadortes inteligentes que, através de um motor de inferência, podem processar perguntas e respostas sobre bases de conhecimento. Uma base de conhecimento pode ser, por exemplo, um conjunto de sentenças do tipo se/então, de maneira que determinadas condições dos dados presentes em um atributo podem levar a imputação de determinados valores em outros.

f) Redes Neurais - Consiste na utilização de abordagens conexionistas para imputação, como redes backpropagation. As redes são normalmente treinadas com os casos

completos, e depois questionadas sobre valores para substiuição nos casos incompletos.

2.2.1 ALGORITMO DOS K-VIZINHOS

O algoritmo dos K-Vizinhos (k-NN), é uma das mais simples e robustas técnicas de Aprendizado de Máquina, regularmente aplicado em tarefas de imputação de valores (SOARES, 2007), onde é classificado como um procedimento de imputação *Hot-Deck*. O algoritmo 1 descreve seu funcionamento (TEKNOMO, 2004):

Input: Quantidade de Vizinhos (K)
Input: Base de Dados para Imputação (BDCI)
Output: Base de Dados de Treino (BDCT)

while *Há registros em BDCI* **do**
    Registro com Valor Ausente (RVA) = Retira de BDCI;
    Cria um Vetor de Distâncias (VD);
    while *Há registros em BDCT* **do**
        Registro Doador (RD) = Retira de BDCT;
        Distancia (D) = Calcula a Distância entre RVA e RD;
        Insere D em VD ordenado;
    end
    Vetor de Vizinhos (VV) = Seleção dos K doadores com menor distância;
    Preenche valor ausente em RVA com a média ou moda dos doadores em VV;
end

Algoritmo 1: Pseudo-Código para o Algoritmo dos K-Vizinhos

Para o cálculo de distância, que estabelece um fator numérico para representação da diferença entre o caso a ser imputado e os casos de exemplo, a distância euclidiana é a mais empregada (SOARES, 2007). Sendo D a distância euclidiana, N o número total de atributos da base, RI o registro a ser imputado e RD um registro doador, têm-se a EQ. 2.2.

$$D = \sqrt{\sum_{i=1}^{N} (RI_i - RD_i)^2} \tag{2.2}$$

Para um exemplo prático, considere a base de exemplo na tabela 2.1.

TAB. 2.1: Base de Dados Preenchida

Id	A	B	C	D
1	2	2	11	8
2	3	3	12	7
3	5	9	15	9

Caso o terceiro atributo do primeiro registro fique ausente, teríamos a disposição da tabela 2.2.

TAB. 2.2: Base de Dados com Valores Ausentes em uma Coluna

Id	A	B	C	D
1	2	X	11	8
2	3	3	12	7
3	5	9	15	9

Neste caso, para imputar o valor ausente, o kNN separa o primeiro registro em um conjuto de imputação, e os registros 2 e 3 em um conjunto de treino. Depois, para o registro a ser imputado, calcula as distâncias euclidianadas para os possíveis doadores:

$$d_{t1,t2} = \sqrt{(2-3)^2 + (11-12)^2 + (8-7)^2} = \sqrt{3} \simeq 1,7;$$
$$d_{t1,t3} = \sqrt{(2-5)^2 + (11-15)^2 + (8-9)^2} = \sqrt{26} \simeq 5.$$

Ao considerar K=1, têm-se como registro mais similar o número 2, de forma que o valor "3", do "Atributo B" deste registro, será empregado para preenchimento no registro 1, um resultado satisfatório, uma vez que o valor encontrado é próximo ao original.

2.2.2 K-VIZINHOS COM SUPORTE A CASOS INCOMPLETOS

Quando a ausência de valores na base é multivariada, é possível que o próprio conjunto de treino apresente valores ausentes, uma vez que não é mais possível realizar uma separação linear entre casos completos e incompletos (VANBUUREN, 2006). Neste caso, o cálculo da distância falha sempre que ocorrem casos doadores com valores ausentes, já que a fórmula

exige a comparação e todos os atributos entre o caso doador e o caso a ser imputado (com exceção, é claro, do atributo a ser imputado).

A fim de lidar com este cenário, Jonsson e Whoolin (JONSSON, 2004) desenvolveram uma variante do algoritmo kNN, aqui denominada kNN-IC. O kNN-IC estabelece que mesmo alguns dos casos incompletos do conjunto de treino podem ser empregados como doadores, bastando que eles apresentem valores preenchidos em todos os atributos que o registro selecionado para imputação também apresente, além de possuir, por razões óbvias, valor preenchido no atributo a ser imputado.

Para fins de exemplo, considere a base de exemplo na tabela 2.3. Se o registro 1 for considerado para imputação no atributo "B", os registros 2 e 3 farão parte do conjunto de treino, uma vez que possuem os mesmos atributos preenchidos que o registro 1, e possuem valor preenchido no atributo a ser imputado. O mesmo não é verdade para os registros 4 e 5, que devem ser desconsiderados como doadores.

TAB. 2.3: Base de Dados com Valores Ausentes em Várias Coluna

Id	A	B	C	D
1	2	X	X	8
2	3	3	X	7
3	5	9	X	9
4	4	8	16	X
5	3	X	11	6

Neste caso, as distâncias calculadas para os casos doadores são:

$$d_{t1,t2} = \sqrt{(2-3)^2 + (8-7)^2} = \sqrt{2} \simeq 1,4;$$
$$d_{t1,t3} = \sqrt{(2-5)^2 + (8-9)^2} = \sqrt{10} \simeq 3,3.$$

Ao considerar K=1, a primeira tupla mais próxima continua a mesma do exemplo anterior, e novamente tem o valor "3", do "Atributo B" empregado para preenchimento no "Atributo B" da tupla 1.

É digno de nota que se não fosse o relaxamento do algoritmo, seria impossível processar qualquer tupla de treino. Ainda assim, é igualmente importante destacar que ao utilizar menos colunas, a distância entre as tuplas mudou consideravelmente, o que traz à tona

a possibilidade de que a redução de atributos para comparação pode assimilar doadores indesejados.

2.3 WORKFLOWS

Um workflow é normalmente definido como a automação de um processo de negócio, em todo, ou em parte, onde diferentes componentes colaboram com o processamento de documentos, informações e atividades, de acordo com um conjunto definido de regras (WFMC, 1999). Workflows computacionais operam sobre Sistemas Gerenciadores de Workflow (WFMS), programas capazes de definir, criar, e gerenciar a execução de definições formais de workflows, utilizando-se de linguagens declarativas, programáticas, ou visuais (BAYENS, 2004; WFMC, 1999).

As aplicações de sistemas de workflow são tão amplas quanto os diferentes processos de negócios encontrados na academia, no mercado e na indústria. São capazes de automatizar, em diferentes níveis, processos de transformação e gerenciamento da informação, linhas de produção industrial e experimentos científicos, entre outros.

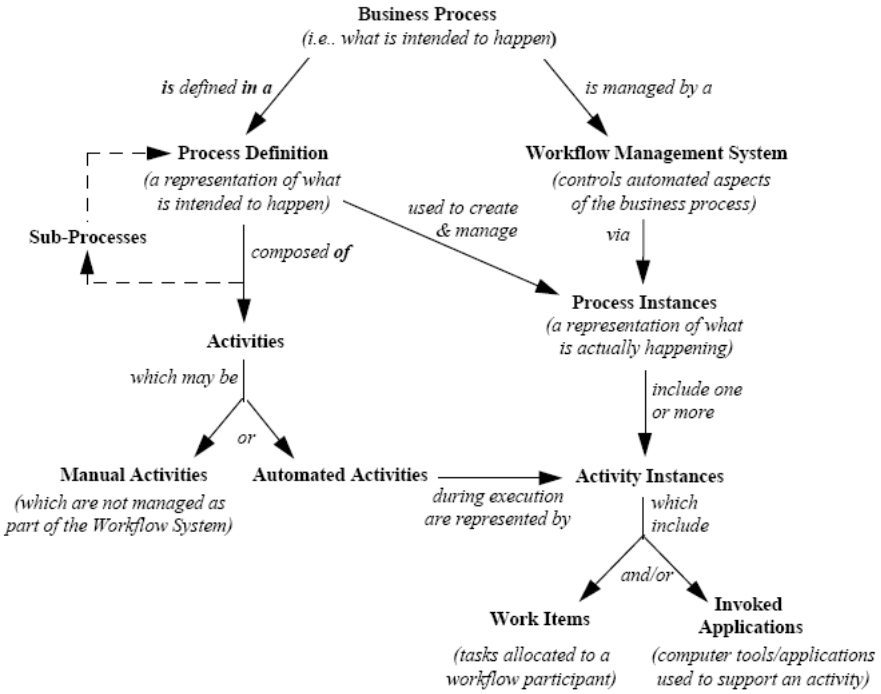


FIG. 2.3: Taxonomia de Workflows - Adaptado de (WFMC, 1999)

A definição de um workflow consiste na estruturação de um processo de negócio, de maneira a suportar sua manipulação automática e computacional. Normalmente, tal

definição é construída como uma cadeia ou grafo de componentes, seus relacionamentos, e marcadores capazes de indicar o começo, a direção de fluxo, e o término do workflow.

A execução de um workflow é a execução de uma definição estruturada por um WFMS, dado um determinado conjunto de informações de entrada e parâmetros específicos de configuração para cada componente. As várias execuções que podem ser derivadas a partir de uma mesma definição são comumente denominadas instâncias de workflow (BAYENS, 2004; WFMC, 1999; HOLLINGSWORTH, 2004; SLOMINSKI, 2003), e resultam da combinação e ajuste-fino dos parâmetros e dados de entrada. A figura 2.3 resume o relacionamento entre os principais conceitos descritos.

3 SOLUÇÃO PROPOSTA

A seção 3.1 descreve em detalhes a metodologia proposta para experimentação em imputação seqüencial, permitindo o reuso de valores imputados. A seção 3.2 apresenta o ambiente desenvolvido para aplicação da metodologia referida.

3.1 METODOLOGIA

Entende-se como metodologia, neste trabalho, um conjunto de práticas, procedimentos e regras, empregados nas atividades de disciplinas específicas (COMPANY, 2004). Neste sentido, a metodologia proposta fornece as bases para a construção de um ambiente configurável, capaz de formalizar e automatizar os passos necessários para a execução de diferentes técnicas de imputação seqüencial, independente do emprego de reuso de valores, como ilustrado na figura 3.1.

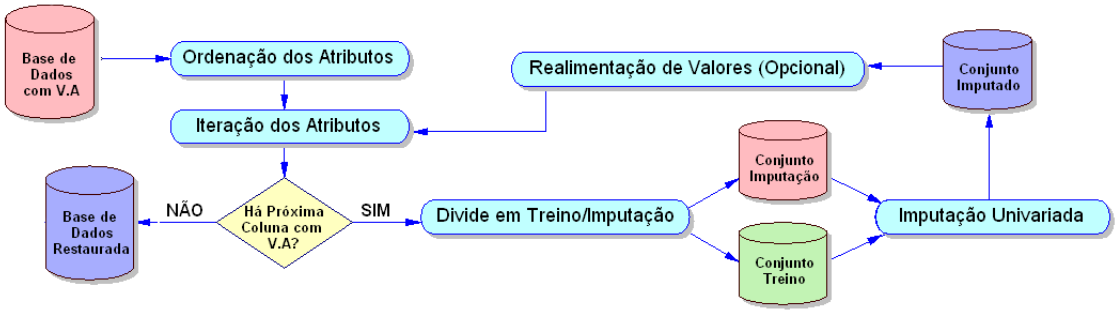


FIG. 3.1: Macro-arquitetura da metodologia proposta

O processo de imputação seqüencial sempre começar a partir de uma base de dados com valores ausentes, sendo a primeira etapa, a ordenação dos atributos. Isto significa determinar uma ordem de prioridade no processamento dos atributos a serem imputados, ou seja, apenas aqueles que possuem valores ausentes. Esta etapa é importante, uma vez que os valores imputados para um atributo podem ser reaproveitados nos conjuntos de treino de imputação do(s) próximo(s), tornando a ordem um fator de impacto capaz de beneficiar ou prejudicar o processo de imputação (IRWIN, 1994). São considerados para este fim apenas os atributos que apresentem valores ausentes, que podem ser ordenados em diferentes critérios, tais como a quantidade de valores ausentes, índices de variância

ou correlação, entropia, entre outros (FERLIN, 2008).

Uma vez ordenados, os atributos são imputados seqüencialmente em um subprocesso iterativo. Para cada atributo, os valores da base de dados são divididos em dois conjuntos, um de treino, e o outro de imputação, sendo o critério de divisão baseado na ocorrência de valor ausente no atributo a ser imputado. Os registros que possuem valor ausente no atributo a ser imputado são armazenados no conjunto de imputação, e os restantes, no de treino.

Após a divisão, é possível executar qualquer algoritmo de imputação univariada, a fim de imputar os valores ausentes para o atributo escolhido, nos registros do conjunto de imputação. É importante destacar que podem ser empregadas quaisquer técnicas de imputação, como a média, regressão linear, redes back-propagation, o algoritmo dos k-vizinhos, entre outros. A aplicação destas técnicas se torna possível devido à característica de divisão-e-conquista da imputação seqüencial, que transforma um problema multivariado em diversos problemas univariados. A imputação univariada dentro deste processo, porém, deve levar em conta alguns fatores novos, como por exemplo, a ocorrência de valores ausentes dentro dos conjuntos de treino, o que leva ao estudo de adaptações dos algoritmos tradicionais. Um exemplo é uma variante do KNN proposta por Jonsson e Woohlin (JONSSON, 2004), empregada neste trabalho.

O resultado da imputação do atributo escolhidos nos registros do conjunto de imputação pode ser inserido na base de dados em uma etapa opcional, caracterizando o reuso de valores. Caso habilitado, o reuso preenche os dados imputados na base de dados antes da iteração seguinte, onde serão aproveitados na formação dos próximos conjuntos de treino. Ao aplicar o reuso de valores, espera-se obter um progressivo preenchimento do conjunto de treino, além de se preservar a consistência entre a imputação de diferentes atributos. Não obstante, valores imputados quase sempre possuem um percentual de erro, e ao reintroduzi-los na base durante o processo de iteração, o analista pode estar, na verdade, contribuindo para uma propagação de erros, ao imputar novos valores sobre dados previamente imputados.

Finalmente, quando não há mais atributos para serem imputados, obtém-se uma base completa. O algoritmo 2 formaliza a metodologia proposta:

Input: Base de Dados com V.A. (BDV)
Output: Base de Dados Restaurada (BDR)

```
Base de Dados Auxiliar (BDX) = Cópia de BDV;
Ordena os atributos de BDX, gerando uma Fila de Atributos a Imputar (FAI);
while Há atributos em FAI do
    Atributo a Imputar (AI) = Retira de FAI;
    Divide BDV em dois Conjuntos ;
      Conjunto de tuplas com Atributo a Imputar ausente (CAA) e;
      Conjunto de tuplas com Atributo a Imputar preenchido (CAC);
    Cria um conjunto vazio de tuplas restauradas (CTR);
    while Há tuplas em CAA do
        Tupla a Imputar (TAI) = Retira de CAA;
        Tupla restaurada (TR) = Imputação Simples(TAI,CAC);
        Inclui TR em CTR;
    end
    Substitui as tuplas de BDX pelas tuplas de CTR;
    if Reuso está habilitado then
        Substitui as tuplas de BDV pelas tuplas de CTR;
    end
end
BDR = BDX;
```

Algoritmo 2: Pseudo-código para o processo de imputação seqüencial

Em suma, a metodologia proposta oferece um meio para a execução de diversas técnicas de imputação seqüencial, e torna o reuso de valores uma técnica independente dos algoritmos utilizados, permitindo a criação e experimentação de diversas combinações diferentes do processo de imputação.

3.2 AMBIENTE DE EXPERIMENTAÇÃO

Com base na metodologia proposta na seção anterior, foi desenvolvido um ambiente cujo principal objetivo é auxiliar na configuração, execução, e análise de experimentos de imputação seqüencial. Sua especificação e implementação são inspiradas nos conceitos fundamentais de workflows, a fim de facilitar a montagem dos componentes, e flexibilizar

a combinação e variação de parâmetros. Foi desenvolvido como uma evolução do software Appraisal, uma aplicação de imputação composta, que combinava algoritmos de seleção e agrupamento na tarefa de imputação de dados, elaborada para utilização na tese de doutorado de Soares (SOARES, 2007).

O ambiente configura automaticamente execuções de workflow que atendam ao processo definido pela metodologia proposta, aliviando o trabalho manual do analista em configurar e re-configurar o processo de imputação a fim de experimentar os melhores parâmetros e combinações de algoritmos (CASTANEDA, 2008).

Ainda, fornece mecanismos para controle e avaliação dos experimentos. É possível para o analista coletar bases originais e completas, para depois provocar artificialmente a ausência de valores e comparar os resultados obtidos por diferentes processos de imputação.

3.2.1 DEFINIÇÕES E INSTÂNCIAS DE WORKFLOW

Ao considerar o processo de imputação definido na metodologia é possível vislumbrar diversas execuções possíveis, seja através da utilização de diferentes algorimos e técnicas para os passos definidos, como das inúmeras combinações possíveis, advindas das variações de parâmetros entre eles. O ambiente proposto procura automatizar a configuração dos experimentos de imputação em ambos os níveis, abordando de maneira bem definida os conceitos de "definição" e "instância" de workflows (WFMC, 1999).

Uma definição consiste em uma possível construção da metodologia, com determinados componentes implementadores de cada passo. Não são explorados na definição de workflow, os parâmetros que estes componentes necessitam para funcionar.

As instâncias de workflow consistem na concretização de uma possível combinação dos parâmetros necessários para os componentes escolhidos em uma definição. Obviamente, uma definição pode possuir diversas instâncias, uma para cada combinação única de todos os parâmetros envolvidos na execução de seus componentes.

A fim de ilustrar melhor os conceitos, a figura 3.2 apresenta apenas a parte do processo relativa ao passo de imputação univariada.

Considerando a figura, cada uma das três etapas pode ser implementada de maneira diferente. Por exemplo, a base de dados carregada pode ser de um SGBD MySQL, ou PostgreSQL, e a imputação de valores pode ser realizada por diversos algoritmos diferentes, como k-NN, redes neurais, média, regressão linear, etc. Ao considerar componentes

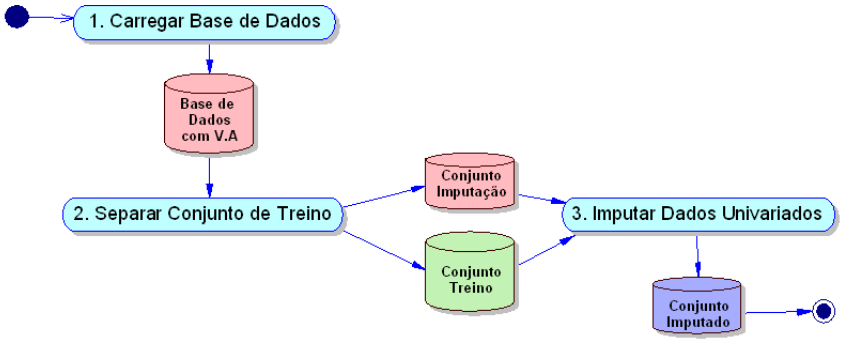


FIG. 3.2: Processo de negócio

concretos para a realização destas etapas, obtém-se a "definição" de um workflow, como ilustrado na figura 3.3:

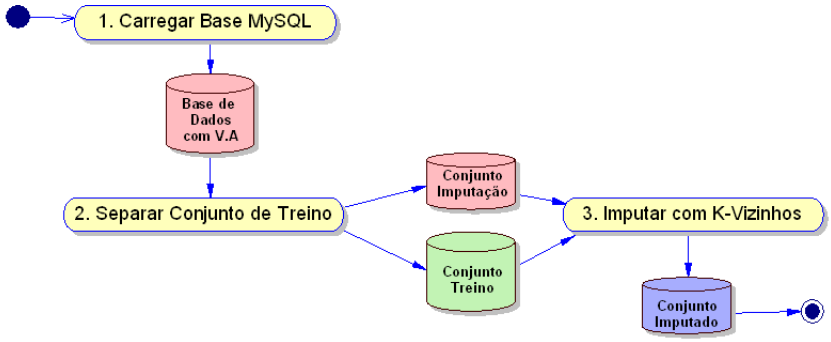


FIG. 3.3: Definição de workflow

Porém, apenas a definição de um workflow ainda não é o suficiente para determinar a maneira como o processo será executado, pois os componentes definidos carecem da configurção de diversos parâmetros. Por exemplo, na imputação com o algoritmo dos K-Vizinhos é possível variar o número de vizinhos e a medida de distância utilizada para o cálculo de vizinhança, um parâmetro conhecido como K.

Ao definir os valores para os parâmetros de configuração de cada componente, obtém-se uma instância de workflow, esta sim, passível de execução por um sistema gerenciador de workflows. Uma possível instância de workflow correspondente ao exemplo é apresentada na figura 3.4.

Cada uma das possíveis instâncias, de cada uma das possíveis definições de workflow possui características próprias, gerando resultados possivelmente diferentes para o processo de imputação em uma mesma base de dados. Assim, é comum que um analista de dados experimente com diversas definições e instâncias a fim de optar por aquela que produza os melhores resultados.

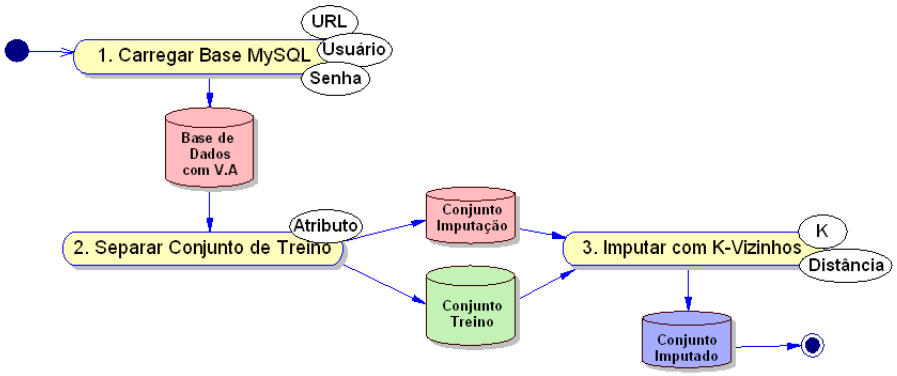


FIG. 3.4: Definição de workflow

3.2.2 REQUISITOS GERAIS DO AMBIENTE

De maneira geral, ao trabalhar separadamente com definições e instâncias, pode-se empregar como componentes quaisquer algoritmos de imputação univariada, ou variações dentro do mesmo algoritmo, permitindo, em outro nível, o ajuste fino de parâmetros e a habilitação de técnicas especiais, como o reuso de valores, ou a normalização de dados.

Os principais requisitos levantados para o desenvolvimento do ambiente de workflow são relacionados com a automação da construção das possíveis combinações de algoritmos e parâmetros, aliviando o trabalho manual de experimentação por parte do analista. As principais funcionalidades para o ambiente são:

- O sistema deve construir e combinar automaticamente diferentes possibilidades de definição do processo de workflow. Isto envolve, por exemplo, detectar diferentes implementações de cada componente, e criar definições individuais para cada combinação única de atividades;

- O sistema deve construir e combinar automaticamente as possíveis instâncias de cada definição de workflow construída. Para tanto, devem ser consideradas todas as variações sobre as entradas de dados dos experimentos, e cada combinação única sobre todos os parâmetros de configuração dos componentes da rede;

- O sistema deve executar automaticamente todas as definições e instâncias construídas, em função de um único comando;

- O sistema deve oferecer recursos de persistência individual para os resultados de cada instância de workflow.

- O sistema deve permitir que os usuários provoquem valores ausentes em bases originais e completas, para serem empregadas em experimentos de imputação.
- O sistema deve oferecer um mecanismo de comparação e avaliação entre os resultados obtidos por diferentes instâncias de workflow.

3.2.3 CONFIGURAÇÃO DO AMBIENTE

A configuração das definições, instâncias, parâmetros de configuração e dados de entrada, pode ser feita através da utilização direta da API de programação do ambiente, ou em arquivos de propriedades, nos quais o usuário determina os componentes habilitados no workflow, e indica a variação dos diversos parâmetros de configuração de cada componente. A figura 3.5 mostra um trecho do arquivo de configuração:

```
# Imputação Simples
#  "knn"      - Knn clássico
#  "bkprop"   - BackPropagation
#  "avg"      - Média
######################
imputation.single.algorithm=knn

# Imputação com KNN
####################
imputation.single.knn.k=1,3,5,10
imputation.single.knn.distance=mahalanobis,euclidian
```

FIG. 3.5: Trecho de configuração do ambiente

No exemplo ilustrado, o usuário pode escolher entre diferentes métodos de imputação univariada, seguindo abaixo, para cada método, os seus respectivos parâmetros de configuração. Como os método de imputação escolhidos foram o algoritmo dos K-Vizinhos e a Média, o ambiente deve construir pelo menos duas definições de workflow, uma cuja estrutura contemple o componente com o algoritmo dos K-Vizinhos e todos os outros componentes definidos, e outra da mesma forma, porém co o componente de imputação com Média. Caso entre os outros componentes também houvesse mais de uma possibilidade de definição, o ambiente resolveria a análise combinatória de todas as possibilidades.

Para cada definição são exploradas as diversas instâncias, pela análise dos valores definidos para os parâmetros de cada componente onde algumas sintaxes são suportadas. Por exemplo, no parâmetro K estão definidos cinco valores separados por vírgula, o que

quer dizer que serão construídas todas as possíveis instâncias com cada um dos valores. É possível ainda definir valores na forma X;Y;Z, onde os valores variam de X a Y, com incremento de Z. Assim, o analista configura diversos valores diferentes para K, e deixa o ambiente explorar todas as possíveis instâncias de criação. Com este mecanismo é possível executar sem esforço, e de uma única vez, diversos experimentos diferentes.

A figura 3.6 ilustra o ambiente exibindo uma definição elaborada para executar a imputação seqüencial de dados.

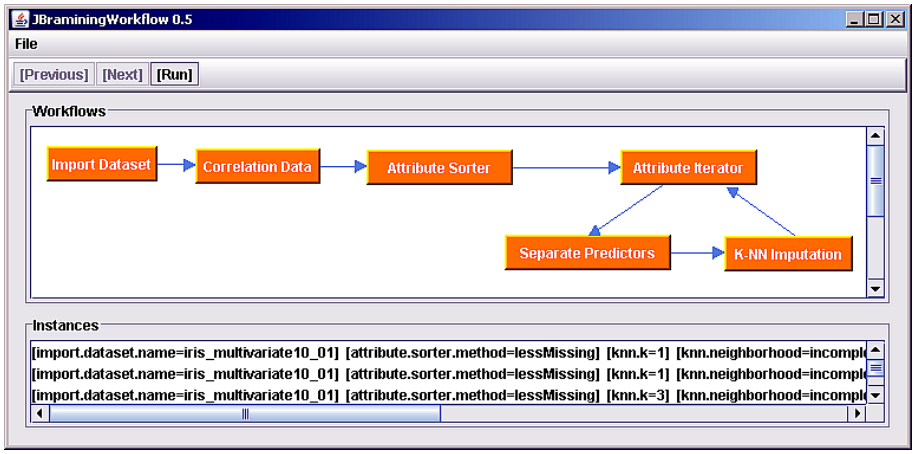


FIG. 3.6: Definição de workflow para imputação

A definição ilustrada encadeia os componentes, de maneira similar ao esquema conceitual apontado na metodologia. Pode-se notar a utilização do algoritmo dos K-Vizinhos para a etapa de imputação. Outras definições poderiam ser montadas utilizando-se por exemplo, BackPropagation, ou a média simples, já que o ambiente é capaz de lidar com mais de uma definição simultaneamente. O usuário pode navegar entre diferentes definições utilizando os botões "previous" e "next". As instâncias podem ser observadas na lista ao final da janela.

A figura 3.7 ilustra a definição que realiza a imputação seqüencial com reuso habilitado. Nesta figura, é possível encontrar o componente opcional, responsável pelo reuso de dados.

Ao clicar em "run", todas as instâncias de todas as definições são automaticamente executadas pelo ambiente.

3.2.4 EXPERIMENTAÇÃO E AVALIAÇÃO

A fim de utilizar o ambiente para a condução e avaliação de experimentos comparativos entre diferentes processos de imputação seqüencial, o analista deve primeiro gerar os

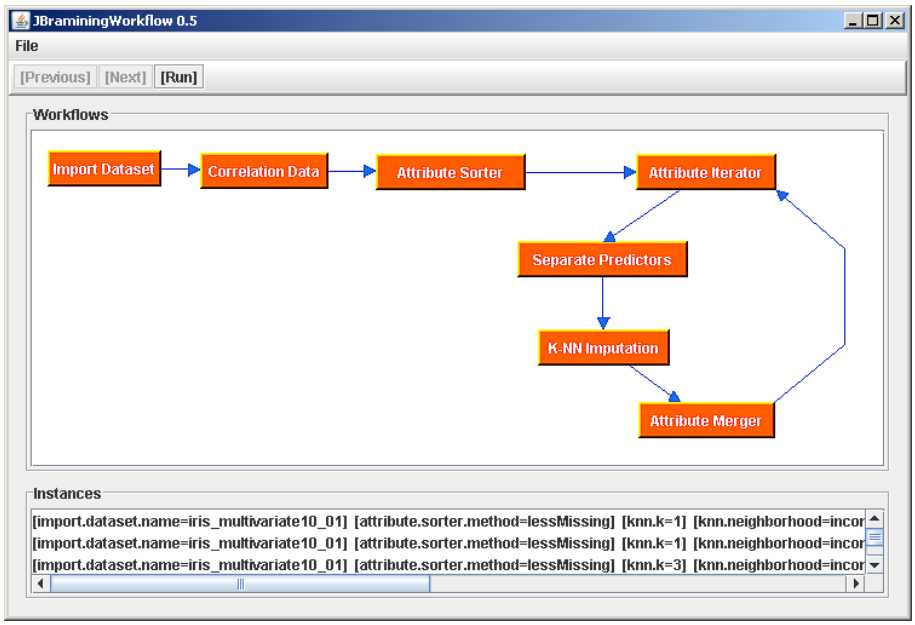


FIG. 3.7: Definição de workflow para imputação com reuso

valores ausentes a partir de uma base original e completa de dados, para depois avaliar os valores resultantes da imputação de cada instância de workflow. A base original é necessária para comparação com os valores gerados pelas instâncias.

3.2.4.1 GERAÇÃO DE VALORES AUSENTES

O módulo de Geração de Valores Ausentes provoca valores ausentes em uma cópia de uma base de dados original e completa. O resultado do seu processamento é uma ou várias bases-problema para serem trabalhadas pelo módulo de imputação seqüencial. Sua utilização é importante, pois os experimentos não podem ser feitos em bases que já possuam valores ausentes, uma vez que não é possível saber o quão melhor ou pior foi um determinado processo de imputação, sem possuir os dados originais para conferência.

Ao provocar os valores ausentes, é possível trabalhar com três mecanismos de ausência: aleatório (MAR - Missing at Random); completamente aleatório (MCAR - Missing Completely at Random); e não-aleatório (NMAR - Not Missing at Random) (LITTLE, 2003).

Em bases com ocorrência de MCAR, os atributos apresentam um padrão completamente aleatório de ausência, sendo impossível precisar qualquer causa para a ocorrência dos valores ausentes. Ao considerar uma base com dados sobre condições atmosféricas, quaisquer atributos poderiam estar ausentes em qualquer registro, independente das medidas.

Quando a ocorrência de um valor ausente está condicionada aos valores encontrados em outros atributos, caracteriza-se a ocorrência de MAR, como por exemplo, caso a umidade do ar apresentasse valores ausentes sempre que a temperatura ambiente fosse menor que 16º.

A diferença do padrão anterior para o NMAR se dá pelo conhecimento da causa que provoca os valores ausentes. Por exemplo, caso os medidores de umidade do ar apresentassem defeito de funcionamento a temperaturas menores que 16º, o mecanismo poderia ser classificado como não-aleatório.

O ambiente oferece, para qualquer mecanismo escolhido, a parametrização do percentual de valores ausentes a ser provocado, bem como os atributos nos quais é permitida a ocorrência de valores ausentes. Em especial, para os mecanismos MAR e NMAR, onde a ausência de valores não é completamente aleatória, apresenta meios para a construção de regras do tipo SE-ENTÃO, que determinam critérios para a marcação dos valores ausentes.

3.2.4.2 AVALIAÇÃO DOS RESULTADOS

Os resultados de imputação são escritos em disco, no formato de XML (para os valores imputados), e Excel (para o resultado das análises). A figura 3.8 mostra um trecho do resultado em XML para os dados de um experimento de imputação.

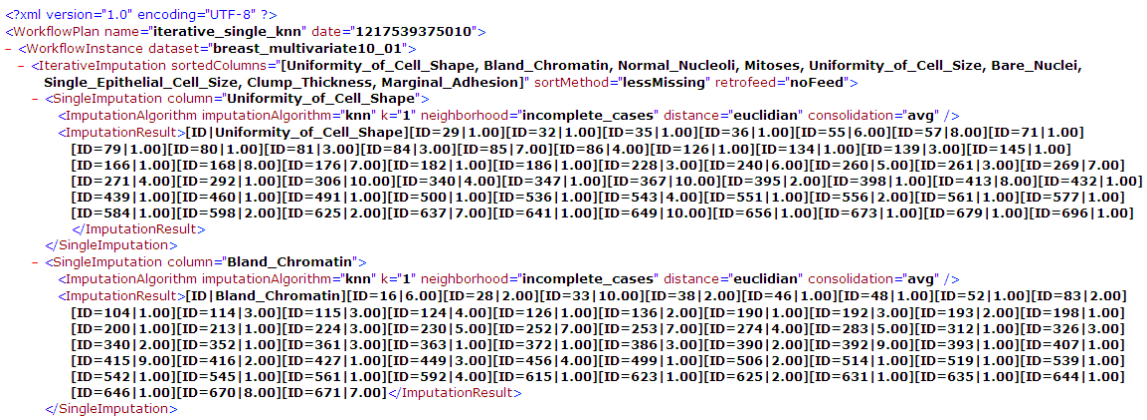


FIG. 3.8: XML de saída do ambiente

O esquema apresenta as definições (*WorkflowPlan*) executados, e para cada plano suas instâncias (*WorkflowInstance*). Fora os valores imputados, todos os parâmetros do contexto do workflow são armazenados, como a ordem das colunas, se houve ou não reuso de dados, o algoritmo utilizado para imputação, entre outros.

O módulo de Análise de Resultados examina os dados imputados, comparando-os com os dados originais, a fim de identificar as diferenças entre os resultados obtidos na execução de diversas instâncias do método proposto. A comparação é feita com base em métricas pré-definidas, sendo que neste trabalho foram empregadas as seguintes:

a) Erro ou distância normalizada (LEPKOWSKI, 2001).

b) Desvio de correlação (FERLIN, 2008).

c) Tempo de execução (GELMAN, 2007).

O conceito de erro, ou distância, é um dos mais empregados na avaliação de técnicas de imputação. A medida mais simples de erro é a diferença absoluta entre o valor original e o valor imputado (Absolute Deviation), de maneira que quanto mais próximo de zero, melhor a qualidade da imputação. Considerando x como o valor ausente de um registro, têm-se a EQ. 3.1.

$$\Delta ad(x_{imputado}, x_{original}) = |x_{imputado} - x_{original}| \tag{3.1}$$

Esta medida de erro porém, não reflete adequadamente a precisão do dado imputado, uma vez que não leva em conta a escala dentro da qual a imputação se encontra. Por exemplo, uma diferença final de 0,3 pode ser um bom resultado para o peso de uma pessoa em quilos, mas um péssimo resultado para a avaliação da espessura da pele, em milímetros.

Para compensar o problema da ordem de grandeza que a variável pode assumir, é comum a utilização do erro relativo (EQ. 3.2), medida de erro que é comumente empregada em experimentos sobre imputação univariada (SOARES, 2007).

$$\Delta rad(x_{imputado}, x_{original}) = \frac{|x_{imputado} - x_{original}|}{x_{original}} \tag{3.2}$$

Porém, quando a ausência de valores se apresenta em diversos atributos, é preciso levar em consideração que cada um pode possuir sua própria ordem de grandeza, e o erro relativo se torna incapaz de apresentar uma comparação fiel entre a precisão que um

mesmo algoritmo atinge para cada atributo (LEPKOWSKI, 2001). Um outro problema desta medida é quando o valor original é zero, uma vez que ele é empregado como divisor.

Para resolver ambos os problemas, pode-se empregar uma medida de erro normalizada (EQ. 3.3), considerando A como o atributo no qual o registro x apresenta valor ausente.

$$\Delta stdd(x_{imputed}, x_{original}) = \frac{100 \times |(x_{imputed} - x_{original})|}{|max_A(x) - min_A(x)|} \tag{3.3}$$

Por exemplo, considere como atributo a nota de um estudante, que pode variar de 0 a 10, e um registro cujo valor original armazenava a nota 7.2. Ao calcular o erro entre o valor original e dois valores hipoteticamente imputados, A=5.0 e B=6.8, os seguintes resultados são obtidos:

$$\Delta stdd(7.2, 5.0) = 100 \times (|7.2 - 5.0|/|10 - 0|) \simeq 22\%;$$
$$\Delta stdd(7.2, 6.8) = 100 \times (|7.2 - 6.8|/|10 - 0|) \simeq 4\%.$$

O erro é expresso em termos de porcentagem, e tende a 100% quando o valor imputado se afasta do valor original, e a 0% quando o valor imputado se aproxima do original. Em outras palavras, menor porcentagem significa melhor imputação.

Esta medida considera que o valor imputado se encontra dentro do alcance dos valores máximo e mínimo do atributo, sendo uma boa técnica para a avaliação de procedimentos de deck-imputation, uma vez que estes inferem as sugestões de imputação através da consulta de casos doadores que são similares ao registro com valor ausente, mantendo sempre a imputação dentro do máximo e mínimo possível.

Esta medida pode ser utilizada a fim de calcular o erro de imputação em um atributo como um todo, através da média dos erros encontrados para cada valor preenchido. Considerando R como o número total de registros com valores ausentes, têm-se a EQ. 3.4.

$$\Delta atributo = \sum_{i=1}^{R} \frac{\Delta stdd(x_{i_{imputed}}, x_{i_{original}})}{R} \tag{3.4}$$

Finalmente, tirando a média entre os erros sumarizados para cada atributo, é possível obter um valor final para a base, que resume a qualidade do processo de imputação em um único indicador, considerando como N o número total de atributos com valores ausentes (EQ. 3.5).

$$\Delta base = \sum_{i=1}^{N} \frac{\Delta atributo_N}{N} \tag{3.5}$$

Outra métrica de erro que pode ser empregada é o desvio de correlação (FERLIN, 2008), que consiste em avaliar se as bases de dados imputadas apresentam uma correlação entre os atributos similar a encontrada na base original.

Esta métrica foi especialmente selecionada para atender a necessidade de avaliação da questão levantada por Schafer (SCHAFER, 1997), de que a imputação seqüencial com o reuso de valores pode aumentar artificialmente a correlação original dos atributos.

A correlação entre dois atributos A e B é dada pela EQ. 3.6, onde σ é o desvio padrão dos valores encontrados no atributo.

$$\rho_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B} \tag{3.6}$$

O resultado final é um número entre -1 e 1, onde -1 significa uma correlação inversamente proporcional perfeita, 0 significa a ausência de correlação, e 1 uma correlação diretamente proporcional perfeita. Ao substituir os valores de um atributo original pelos de um atributo imputado, pode ocorrer um aumento da correlação (i.e. se aproxima de 1 ou -1), ou um enfraquecimento da correlação (i.e. se aproxima de 0).

Para avaliar a correlação completa de uma base emprega-se uma matriz de correlação, que pode ser definida como uma matriz M de tamanho NxN, onde N é o número total de atributos na base. A matriz cobre a correlação entre todos os possíveis pares de atributos (7).

$$\mathbf{M} = \begin{bmatrix} \rho(K_1,K_1) & \rho(K_1,K_2) & \dots & \rho(K_1,K_N) \\ \rho(K_2,K_1) & \rho(K_2,K_2) & \dots & \rho(K_1,K_N) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(K_N,K_1) & \rho(K_N,K_2) & \dots & \rho(K_N,K_N) \end{bmatrix} \tag{3.7}$$

Dentro da matriz, pode-se considerar como a correlação geral de um único atributo, a média da correlação entre este atributo e todos os outros atributos da base. Por exemplo, a correlação geral do atributo K_1 é dada por EQ. 3.8.

$$CGA(K_1) = \frac{\rho(K_1, K_2) + \rho(K_1, K_3) + \cdots + \rho(K1, K_N)}{N - 1} \tag{3.8}$$

A métrica do desvio de correlação é dada, para um atributo, pela diferença entre a correlação geral dos atributos da base imputada, e a correlação geral dos atributos da base original (EQ. 3.9).

$$DA(K_{imputed}, K_{original}) = CGA(K_{imputed}) - CGA(K_{original}) \tag{3.9}$$

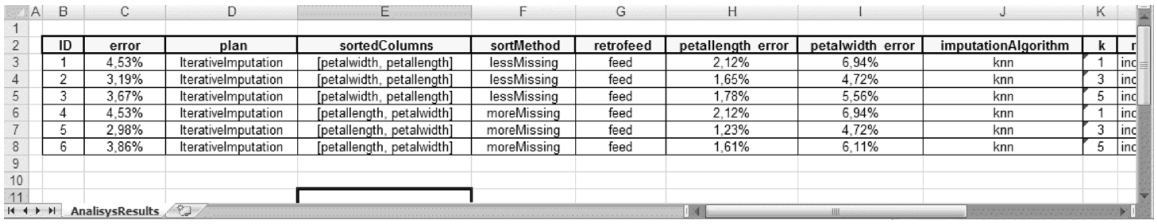
Esta fórmula gera um valor positivo, caso a correlação final dos dados imputados seja maior que a correlação original (ou seja, se a correlação dos valores imputados se aproximou de 1 ou -1), ou negativo, caso contrário (ou seja, se a correlação dos valores impuatados se aproximou de 0).

Finalmente, somando o desvio de correlação de todos os atributos é possível quantificar em um índice único, o impacto do procedimento de imputação sobre o coeficiente de correlação dos dados originais (EQ. 3.10).

$$\sum_{i=1}^{N} DA(K_{i_{imputed}}, K_{i_{original}}) \tag{3.10}$$

O resultado do processamento da análise é uma planilha similar a ilustrada na figura 3.9.

A planilha mostra, para cada instância, os parâmetros de configuração utilizados, e os resultados obtidos para todas as métricas definidas pelo analista. As métricas, quando possível, são exibidas em suas formas sumarizadas, para toda a base, e de forma individual para cada atributo imputado no processo seqüencial. É possível alterar a configuração



ID	error	plan	sortedColumns	sortMethod	retrofeed	petallength error	petalwidth error	imputationAlgorithm	k	r
1	4,53%	IterativeImputation	[petalwidth, petallength]	lessMissing	feed	2,12%	6,94%	knn	1	inc
2	3,19%	IterativeImputation	[petalwidth, petallength]	lessMissing	feed	1,65%	4,72%	knn	3	inc
3	3,67%	IterativeImputation	[petalwidth, petallength]	lessMissing	feed	1,78%	5,56%	knn	5	inc
4	4,53%	IterativeImputation	[petallength, petalwidth]	moreMissing	feed	2,12%	6,94%	knn	1	inc
5	2,98%	IterativeImputation	[petallength, petalwidth]	moreMissing	feed	1,23%	4,72%	knn	3	inc
6	3,86%	IterativeImputation	[petallength, petalwidth]	moreMissing	feed	1,61%	6,11%	knn	5	inc

AnalisysResults

FIG. 3.9: Planilha de análise gerada pelo ambiente

do ambiente para que os resultados sejam exportados também para bancos de dados relacionais, como MySQL ou PostgreSQL, além de arquivos de texto separados por vírgula (CSV).

3.2.5 CRIAÇÃO DE NOVOS COMPONENTES

O ambiente proposto procura facilitar a integração de novos algoritmos de imputação, ou tarefas de uso geral. Criar um novo componente significa desenvolver uma classe cuja única restrição é a necessidade de se implementar uma interface provida pela API, que determina uma mesma assinatura para o método de invocação de todos os componentes, uma solução baseada no padrão de projeto "Command" (GAMMA, 1995). A figura 3.10 ilustra o conceito:

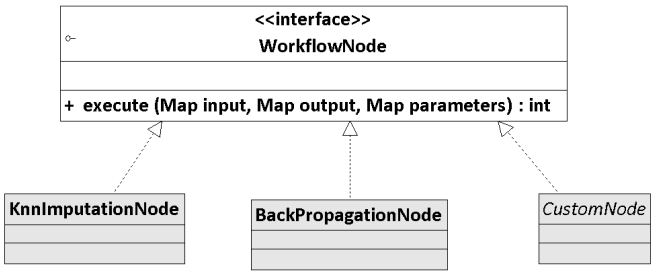


FIG. 3.10: Interface a ser implementada pelos componentes

Isto significa que a invocação das tarefas do workflow é realizada de maneira uniforme, porém, a maneira como a tarefa é processada fica a cargo de cada componente, cujas classes podem fornecer implementações próprias, ou atuar como fachadas para invocação de componentes prontos ou sistemas legados.

É necessário, além de desenvolver as classes, especificar os parâmetros esperados pelo componente, para que o ambiente possa criar corretamente as instâncias de workflow. Tal especificação é feita em um descritor XML interno.

Utilizando esta estratégia já foi possível integrar ao ambiente componentes que utilizam, por exemplo, redes neurais do framework Java Object Oriented Neural Networks (JOONE) (MARRONE, 2007), algoritmos evolucionários do Java Genetic Algorithms Package (JGAP) (N., 2007) e do Java Particle-Swarm Optimization Package (JSwarm-PSO) (CINGOLANI, 2005), entre outros.

4 EXPERIMENTOS E RESULTADOS

Este capítulo detalha o escopo dos experimentos realizados e os resultados encontrados. Apresenta ainda uma análise comparativa dos resultados, por base de dados, e em geral.

4.1 ESCOPO DOS EXPERIMENTOS

4.1.1 BASES DE DADOS

Foram utilizadas nos experimentos cinco bases de dados, provenientes do repositório de bases para Machine Learning e KDD da Universidade da Califórnia, em Irvine (MERZ, 1998): Iris Plants, Pima Indians Diabetes, Wiscosin Breast Cancer, Computer Hardware e Wine.

Apesar das bases de dados selecionadas serem de diferentes domínios, esta dissertação trata apenas dados numéricos, de maneira que, desconsiderando-se as colunas determinantes de classe, todas apresentam atributos contínuos. A única exceção é Computer Hardware, que apesar de possuir atributos numéricos, como quantidade de memória, cache e clock de CPU, contém valores discretos, já que os modelos de computadores e de dispositivos de hardware são pré-estabelecidos pelo mercado.

As subseções a seguir detalham cada uma das bases, apresentando, entre outras informações, o número de atributos e registros, a média, o desvio padrão, e a correlação entre os atributos.

TAB. 4.1: Atributos e registros dos conjuntos de dados

Base	**Atributos**	**Tuplas**
Iris Plants	4	150
Pima Indians Diabetes	8	768
Breast Cancer	10	699
Computer Hardware	9	209
Wine	13	178

4.1.1.1 IRIS PLANTS

A base Iris Plants é uma das mais utilizadas pelos pesquisadores, e consta em praticamente todos os trabalhos relacionados. Contém informações sobre as medidas de comprimento e largura das pétalas e caule de plantas de três espécies: Virginica, Versicolor e Setosa. Das 150 tuplas na base, há 50 de cada espécie. É conhecida por ser uma base bem comportada, com poucos atributos, ordens de grandeza similares, bom grau de correlação, além da perfeita distribuição de classes.

A tabela 4.2 apresenta um sumário dos atributos, e a 4.3 a matriz de correlação.

TAB. 4.2: Iris Dataset - descrição dos atributos

Atributo	Unidade	Val Mínimo	Val Máximo	Média	D Padrão
sepallength	Real	4.3	7.9	5.84	0.83
sepalwidth	Real	2.0	4.4	3.05	0.43
petallength	Real	1.0	6.9	3.76	1.76
petalwidth	Real	0.1	2.5	1.20	0.76

TAB. 4.3: Iris Dataset - correlação dos atributos

	sepallength	sepalwidth	petallength	petalwidth
sepallength	1.00	-0.11	0.82	0.87
sepalwidth	-0.11	1.00	-0.36	-0.42
petallength	0.87	-0.42	1.00	0.96
petalwidth	0.82	-0.36	0.96	1.00

Observando a matriz de correlação é possível constatar que as colunas *petallength*, *petalwidth* e *sepallength* possuem um alto grau de correlação entre si, sendo a mais alta entre *petallength* e *petalwidth*. Isto que dizer que, por exemplo, para as pétalas das plantas da base, a largura possui relação com o comprimento, porém o mesmo não acontece com os caules.

4.1.1.2 PIMA INDIANS DIABETES

A base Pima Indians Diabetes relaciona dados sobre o diagnóstico de diabetes em mulheres de uma tribo indígena. Possui 768 registros, sendo 500 com casos negativos, e 268 com resultado positivo para esta doença.

A tabela 4.4 apresenta um sumário dos atributos, e a 4.5 a matriz de correlação.

TAB. 4.4: Pima Indians Dataset - descrição dos atributos

Atributo	Tipo	Val. Mínimo	Val. Máximo	Média	D. Padrão
Pedigree function	real	0,085	2,42	0,52	0,34
Glucose concentration	int	56	198	122,62	30,82
Body mass	real	18,2	67,1	33,08	7,01
Skin fold thickness	int	7	63	29,14	10,50
Blood Pressure	int	24	110	70,66	12,48
Age	int	21	81	30,86	10,18
Serum insulin	int	14	846	156,05	118,69
Pregnancy times	int	0	17	3,30	3,20

TAB. 4.5: Pima Indians Dataset - correlação dos atributos

	PF	GC	BM	SK	BP	AG	SI	PT
PF	1.00	0.14	0.16	0.16	-0.02	0.09	0.14	0.01
GC	0.14	1.00	0.21	0.20	0.21	0.34	0.58	0.20
BM	0.16	0.21	1.00	0.66	0.30	0.07	0.23	-0.03
SK	0.16	0.20	0.66	1.00	0.23	0.17	0.18	0.09
BP	-0.02	0.21	0.30	0.23	1.00	0.30	0.10	0.21
AG	0.09	0.34	0.07	0.17	0.30	1.00	0.22	0.68
SI	0.14	0.58	0.23	0.18	0.10	0.22	1.00	0.08
PT	0.01	0.20	-0.03	0.09	0.21	0.68	0.08	1.00

Apesar de ser descrita como uma base completamente preenchida pelo repositório da UCI, é possível constatar a presença de valores inconsistentes na Pima, nos atributos *blood_pressure*, *body_mass*, *glucose_concentration skin_fold_thickness* e *serum_insulin*. No total, existem 376 registros onde um ou mais destes atributos apresentam valor 0, o que não deveria ser possível, uma vez que não existem pessoas, por exemplo, sem pressão sanguínea ou massa corporal. Sendo assim, estes registros foram removidos do conjunto original de dados, ficando a base com 392 registros, sendo 262 de casos positivos, e 130 negativos. Esta ocorrência já foi observada em outros trabalhos de imputação, tendo-se tomado a mesma decisão em remover os registros (FERLIN, 2008; SOARES, 2007).

A Pima é a base que apresenta os menores índices de correlação, todos muito próximos de zero. Os maiores índices observados estão entre os atributos idade e número de gravidez (*age* e *pregnancy_times*) e massa corporal e espessura da pele (*body_mass* e *skin_fold_thickness*).

4.1.1.3 WISCOSIN BREAST CANCER

A Wiscosin Breast Cancer é uma base de dados do hospital da Universidade de Wisconsin, que armazena informações sobre diagnósticos de câncer de mama. São 682 registros completos (de um total de 699), com 239 pacientes com diagnóstico positivo, e 443 que não apresentam a ocorrência do câncer.

A tabela 4.6 apresenta um sumário dos atributos, e a 4.7 a matriz de correlação.

TAB. 4.6: Breast Cancer Dataset - descrição dos atributos

Atributo	Tipo	Val. Mínimo	Val. Máximo	Média	D. Padrão
Uniformity of Cell Size	int	1	10	3,15	3,06
Clump Thickness	int	1	10	4,43	2,82
Bland Chromatin	int	1	10	3,44	2,44
Uniformity of Cell Shape	int	1	10	3,21	2,98
Marginal Adhesion	int	1	10	2,83	2,86
Mitoses	int	1	10	1,60	1,73
Bare Nuclei	int	1	10	3,54	3,64
Normal Nucleoli	int	1	10	2,87	3,05
Single Epithelial Cell Size	int	1	10	3,23	2,22

TAB. 4.7: Breast Cancer Dataset - correlação dos atributos

	UCSZ	CT	BC	UCSH	MA	MIT	BN	NN	SECS
UCSZ	1.00	0.64	0.76	0.91	0.71	0.46	0.69	0.72	0.75
CT	0.64	1.00	0.55	0.65	0.49	0.35	0.59	0.53	0.52
BC	0.76	0.55	1.00	0.74	0.67	0.35	0.68	0.67	0.62
UCSH	0.91	0.65	0.74	1.00	0.69	0.44	0.71	0.72	0.72
MA	0.71	0.49	0.67	0.69	1.00	0.42	0.67	0.60	0.59
MIT	0.46	0.35	0.35	0.44	0.42	1.00	0.34	0.43	0.48
BN	0.69	0.59	0.68	0.71	0.67	0.34	1.00	0.58	0.59
NN	0.72	0.53	0.67	0.72	0.60	0.43	0.58	1.00	0.63
SECS	0.75	0.52	0.62	0.72	0.59	0.48	0.59	0.63	1.00

A base apresenta uma alta correlação entre os seus atributos, com exceção de Mitoses, que possui índices abaixo de 0,5 com as outras colunas. A maior correlação acontece entre a uniformidade do tamanho da célula e a uniformidade do formato da célula (*Uniformity_of_Cell_Size* e *Uniformity_of_Cell_Shape*).

4.1.1.4 COMPUTER HARDWARE

A base Computer Hardware contém dados do desempenho de microcomputadores. Relaciona informações como tamanho da memória, clock do processador, memória cache, entre outros, não possui porém, um atributo de classe. Contém 209 tuplas e nenhuma apresenta valores ausentes.

Um fator importante sobre esta base é que os dados dos atributos, apesar de numéricos, se apresentam em um domínio discreto, uma vez que os possíveis valores para as unidades de hardware dos microcomputadores são determinados pelos fabricantes.

Por exemplo, um pente de memória possui apenas valores de base 2, como 128Mb, 256Mb ou 1024Mb. Ainda assim, é possível que em um computador sejam instaladas combinações incomuns destes pentes, levando a configurações que não são normalmentes encontradas. Estas características tornam este domínio bastante peculiar e interessante para o problema de imputação de valores.

A tabela 4.8 apresenta um sumário dos atributos, e a 4.9 a matriz de correlação:

TAB. 4.8: Computer Hardware Dataset - descrição dos atributos

Atributo	Tipo	Val. Mínimo	Val. Máximo	Média	D. Padrão
MemMax	int	64	64.000	11.796,1	11.726,6
MemMin	int	64	32.000	2.868	3.878,7
ChannelsMax	int	0	176	18,2	26,0
ChannelsMin	int	0	52	4,7	6,8
Estimated Rel Perf	int	15	1.238	99,3	154,8
MachineCycleTime	int	17	1.500	203,8	260,3
Public Rel Perf	int	6	1.150	105,6	160,8
Cache Memory	int	0	256	25,2	40,6

TAB. 4.9: Computer Hardware Dataset - correlação dos atributos

	MMAX	MMIN	CMAX	CMIN	ERP	MCT	PRP	CM
MMAX	1.00	0.76	0.53	0.56	0.90	-0.38	0.86	0.54
MMIN	0.76	1.00	0.27	0.52	0.82	-0.34	0.79	0.53
CMAX	0.53	0.27	1.00	0.55	0.59	-0.25	0.61	0.49
CMIN	0.56	0.52	0.55	1.00	0.61	-0.30	0.61	0.58
ERP	0.90	0.82	0.59	0.61	1.00	-0.29	0.97	0.65
MCT	-0.38	-0.34	-0.25	-0.30	-0.29	1.00	-0.31	-0.32
PRP	0.86	0.79	0.61	0.61	0.97	-0.31	1.00	0.66
CM	0.54	0.53	0.49	0.58	0.65	-0.32	0.66	1.00

Os atributos da computer hardware possuem um alto fator de correlação direta, o que corresponde a expectativa comum dentro dos microcomputadores, ou seja, um microcomputador com mais memória principal, normalmente possui mais memória cache. O único caso contrário é o tempo de ciclo de CPU (*cycle-time*), que possui um alto fator de correlação inversa (correlação negativa), o que quer dizer que quanto maior o desempenho da máquina, menor o tempo de ciclo da CPU.

4.1.1.5 WINE

A base Wine contém dados químicos sobre amostras de vinhos, em uma análise feita sobre treze atributos comumente encontrados em três tipos de vinho diferentes. São 178 registros, sendo 59 do primeiro tipo, 71 do segundo, e 48 do terceiro.

A tabela 4.10 apresenta um sumário dos atributos, e a 4.11 a matriz de correlação:

TAB. 4.10: Wine Dataset - descrição dos atributos

Atributo	Tipo	Val. Mínimo	Val. Máximo	Média	D. Padrão
Proanthocyanins	real	0,41	3,58	1,75	0,54
Magnesium	int	70	162	100,72	15,39
Hue	real	0,70	1,71	1,07	0,16
Nonflavanoid phenols	real	0,13	0,66	0,33	0,11
Ash	real	1,36	3,23	2,34	0,30
Alcalinity of ash	real	10,60	30	18,65	3,29
Total phenols	real	1,10	3,88	2,53	0,56
Proline	int	278	1680	811,66	349,77
Alcohol	real	11,03	14,83	12,98	0,89
Color intensity	real	1,28	8,90	4,27	1,63
Flavanoids	real	0,57	5,08	2,49	0,75
OD280/OD315	real	1,59	4,00	2,95	0,48
Malic acid	real	0,74	4,43	1,90	0,78

Na base Wine os atributos possuem, de maneira geral, um baixo fator de correlação, inclusive com casos onde o índice é praticamente zero (sem qualquer correlação). A exceção fica por conta dos atributos *Proantho Cyaning*, *Flavanoids*, e *OD280/OD315*, que possuem um índice de correlação maior que 0,6.

4.1.2 AUSÊNCIA DE VALORES

A partir das bases de dados originais, com o módulo Eraser do ambiente de imputação, foram construídas diversas versões com valores ausentes. O mecanismo de ausência em-

TAB. 4.11: Wine Dataset - correlação dos atributos

	PROA	MGS	HUE	NFL	ASH	ALC	TPH	PROL	ALCO	CI	FLA	OD2	MA
PROA	1.00	0.24	0.30	-0.37	0.01	-0.20	0.61	0.33	0.14	-0.03	0.65	0.52	-0.22
MGS	0.24	1.00	0.06	-0.26	0.29	-0.08	0.21	0.39	0.27	0.20	0.20	0.07	-0.05
HUE	0.30	0.06	1.00	-0.26	-0.07	-0.27	0.43	0.24	-0.07	-0.52	0.54	0.57	-0.56
NFL	-0.37	-0.26	-0.26	1.00	0.19	0.36	-0.45	-0.31	-0.16	0.14	-0.54	-0.50	0.29
ASH	0.01	0.29	-0.07	0.19	1.00	0.44	0.13	0.22	0.21	0.26	0.12	0.00	0.16
ALC	-0.20	-0.08	-0.27	0.36	0.44	1.00	-0.32	-0.44	-0.31	0.02	-0.35	-0.28	0.29
TPH	0.61	0.21	0.43	-0.45	0.13	-0.32	1.00	0.50	0.29	-0.06	0.86	0.70	-0.34
PROL	0.33	0.39	0.24	-0.31	0.22	-0.44	0.50	1.00	0.64	0.32	0.49	0.31	-0.19
ALCO	0.14	0.27	-0.07	-0.16	0.21	-0.31	0.29	0.64	1.00	0.55	0.24	0.07	0.09
CI	-0.03	0.20	-0.52	0.14	0.26	0.02	-0.06	0.32	0.55	1.00	-0.17	-0.43	0.25
FLA	0.65	0.20	0.54	-0.54	0.12	-0.35	0.86	0.49	0.24	-0.17	1.00	0.79	-0.41
OD2	0.52	0.07	0.57	-0.50	0.00	-0.28	0.70	0.31	0.07	-0.43	0.79	1.00	-0.37
MA	-0.22	-0.05	-0.56	0.29	0.16	0.29	-0.34	-0.19	0.09	0.25	-0.41	-0.37	1.00

pregado foi o MCAR, em três percentuais de ausência, com respectivamente 10%, 20% e 30% de todas as células de cada base original. Estes valores foram escolhidos por sua referência na literatura (FERLIN, 2008; SOARES, 2007), e através de testes empíricos, onde constatou-se que a imputação de valores a partir de 40% das células era impraticável. Para efeitos de amostragem, foram construídas três versões de cada percentual de ausência, para cada base de dados original.

4.1.3 ALGORITMO DE IMPUTAÇÃO

O algoritmo de imputação escolhido para os experimentos foi uma variação do algoritmo dos K-Vizinhos (KNN), proposta por Jonsson e Woolin, especialmente para imputação em cenários de ausência multivariada de valores (JONSSON, 2004). Ele foi escolhido por ser uma técnica tradicional, reconhecidamente robusta e eficiente (SOARES, 2007).

O algoritmo foi utilizado em todos os experimentos, com e sem reutilização de valores, e o parâmetro de configuração K, que determina o número de vizinhos, foi variado em 1, 3, 5 e 10. Muito embora não existam heurísticas reconhecidamente estabelecidas, estes valores são comumente empregados em experimentos de imputação com o algoritmo dos k-vizinhos (FERLIN, 2008; SOARES, 2007).

4.1.4 REUSO DE VALORES

Os experimentos empregaram a imputação seqüencial atributo-a-atributo com e sem reuso. Para os experimentos com reuso, foram empregados os seguintes critérios de ordenação (FERLIN, 2008):

- Do atributo com menos valores ausentes para o atributo com mais valores ausentes.

- Do atributo com mais valores ausentes para o atributo com menos valores ausentes.
- Do atributo com menor correlação para o atributo com maior correlação.
- Do atributo com maior correlação para o atributo com menor correlação.
- Sem ordenação, estabelecendo-se a ordem proveniente das consultas ao banco.

4.1.5 SUMÁRIO DOS EXPERIMENTOS

Para execução dos experimentos propostos foram estabelecidas duas definições de workflow para imputação sequencial atributo-a-atributo, uma com o componente de reuso de valores habilitado, e a outra sem.

A tabela 4.12 ilustra as diferentes combinações de parâmetros, e o número total de instâncias configuradas, executadas, e avalidas pelo ambiente de imputação, para cada uma das duas definições.

TAB. 4.12: Definições e instâncias do experimento

Definição 01 - Imp. Sequencial sem Reuso - 12 Instâncias		
Mec. de Ausência	Perc. de Ausência	Knn - K
MAR	10%,20%,30%	1,3,5,10

Definição 02 - Imp. Sequencial com Reuso - 60 Instâncias			
Mec. de Ausência	Perc. de Ausência	Ordem de Imputação	Knn - Nº de Vizinhos
MAR	10%,20%,30%	Sem ordem, Menos V.A, Mais V.A, Menor Corr., Maior Corr.	1,3,5,10

Estas configurações foram empregadas em cada base de dados utilizada nos experimentos. Considerando que foram empregadas cinco bases, com cinco versões de cada uma, o total é de 1800 experimentos realizados, cuja execução tomou cerca de 70 horas de processamento, em máquinas Pentium IV com 512Mb de memória principal.

4.2 ANÁLISE DOS RESULTADOS

Os resultados são apresentados para cada uma das bases utilizadas nos experimentos, com gráficos de linha para a imputação com e sem reuso de valores, para representar a evolução do erro de imputação e do desvio de correlação em função do percentual de ausência.

Foram escolhidos para representação nos gráficos as melhores instâncias (i.e. menor erro de imputação) de cada definição, respectivamente com e sem reuso de valores.

4.2.1 IRIS PLANTS DATASET

Na base Iris, o reuso de valores não melhora significativamente o processo de imputação, e seu desempenho praticamente iguala o dos experimentos sem reuso de valores, como pode ser visto na figura 4.1. Iris é uma base conhecida como padrão para avaliação de métodos de KDD, combinando uma perfeita distribuição de classes com uma variação muito pequena de valores em cada atributo. Como estas caracerísticas já favorecem qualquer técnica de imputação, é possível que o reuso não tenha sido capaz de contribuir com o processo como um todo.

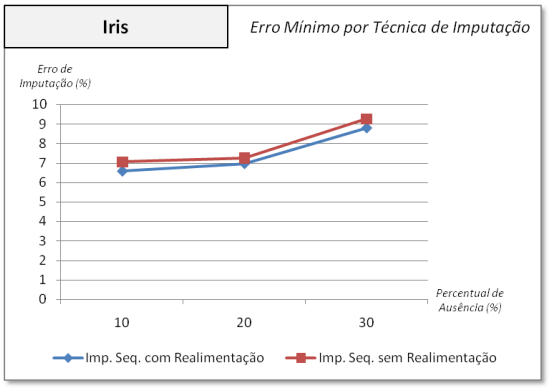


FIG. 4.1: Iris Dataset - erro de imputação

Quanto à avaliação do desvio de correlação, que pode ser observado na figura 4.2, é possível observar que ao habilitar o reuso de valores, a distorção é menor.

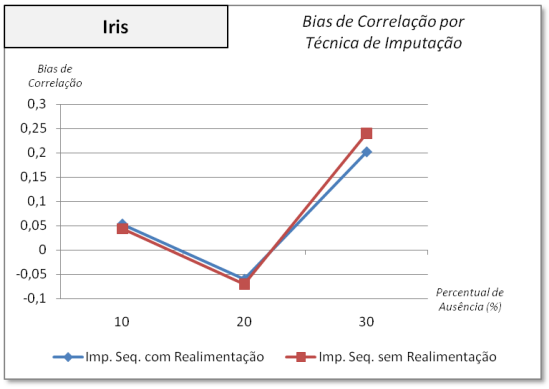


FIG. 4.2: Iris Dataset - desvio de correlação

4.2.2 BREAST CANCER DATASET

Na figura 4.3, que apresenta o erro de imputação para a Breast Cancer dataset, é possível observar que o reuso de valores possui desempenho diferenciado quando o nível percentual de ausência é de 30%. Os atributos na base são altamente correlacionados, de maneira que o reuso pode desempenhar um papel significativo na manutenção da consistência entre os valores imputados.

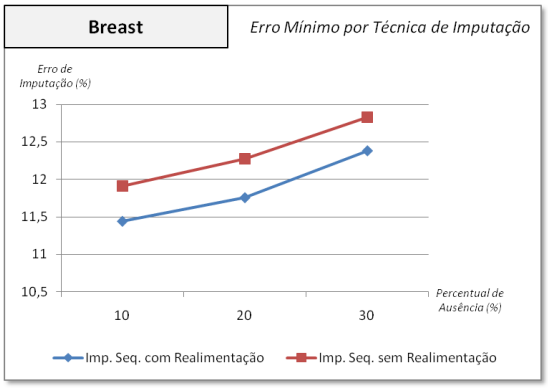


FIG. 4.3: Breast Dataset - erro de imputação

Ao realimentar os valores, obteve-se menores índices de desvio de correlação quando o percentual de ausência atingiu 30%, tornando a Breast Dataset um caso onde habilitar a reutilização é vantajosa tanto na avaliação do erro quanto no desvio de correlação. A figura 4.4 ilustra os índices:

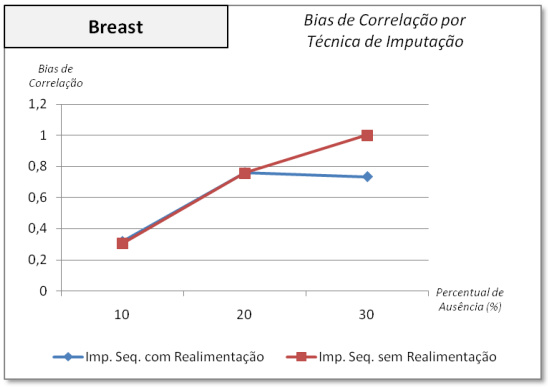


FIG. 4.4: Breast Dataset - desvio de correlação

4.2.3 PIMA DATASET

A Pima é uma base desafiadora para os processos de imputação, uma vez que seus atributos possuem um baixo grau de correlação. Assim, a habilitação do reuso não apresentou melhorias significativas no processo, algumas vezes até piorando a qualidade final do processo de imputação, como quando o percentual de ausência atinge 30% (figura 4.5).

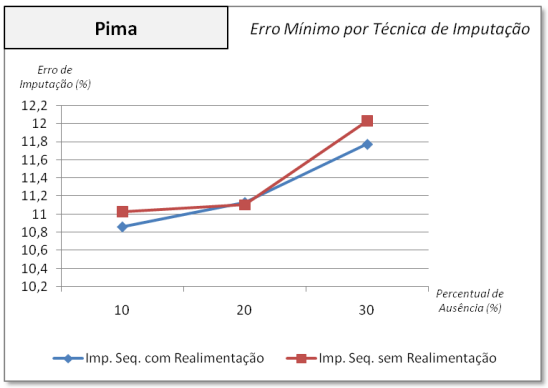


FIG. 4.5: Pima Dataset - erro de imputação

Sobre o desvio de correlação, porém, o reuso de valores teve um impacto significativamente menor sobre o fator original. Como pode ser visto na figura 4.6, enquanto a implementação simples demonstra um crescimento do desvio de correlação, a reutilização manteve os índices mais próximos de zero.

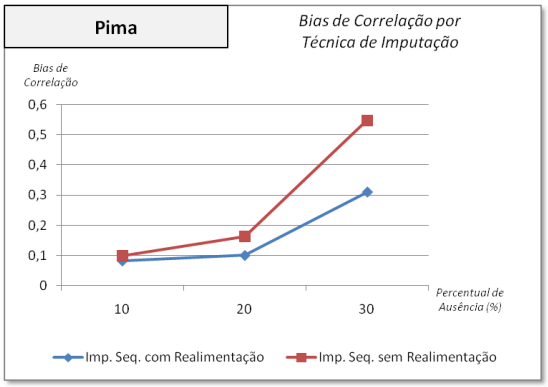


FIG. 4.6: Pima Dataset - desvio de correlação

4.2.4 COMPUTER HARDWARE DATASET

A base Computer Hardware apresenta alguns atributos extremamente correlacionados, o que abriu espaço para que o reuso de valores melhorasse qualidade do processo de

imputação em todos os níveis de ausência, como pode ser observado na figura 4.7.

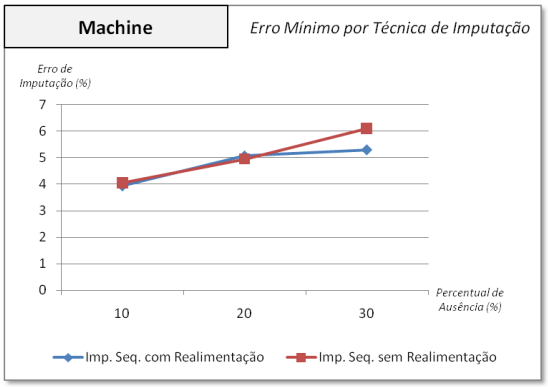


FIG. 4.7: Computer Hardware Dataset - erro de imputação

Por outro lado, a correlação que era bem alta, foi enfraquecida pelo reuso de valores, enquanto a abordagem sem reutilização manteve os índices mais próximos de zero. A diferença no desvio de correlação pode ser observada na figura 4.8.

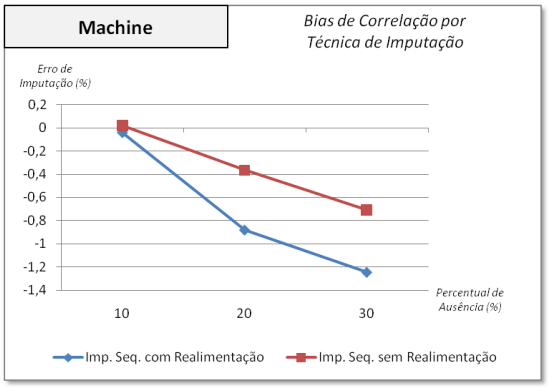


FIG. 4.8: Computer Hardware Dataset - desvio de correlação

4.2.5 WINE DATASET

Na base Wine, o reuso de valores mostrou um impacto constante no processo de imputação, sempre obtendo menores taxas de erro que a implementação sem reuso. O gráfico com os erros é ilustrado na figura 4.9.

Ambas as abordagens provocaram um enfraquecimento na correlação dos atributos, sendo este mais significativo quando habilitando o reuso de valores, como ilustrado na figura 4.10.

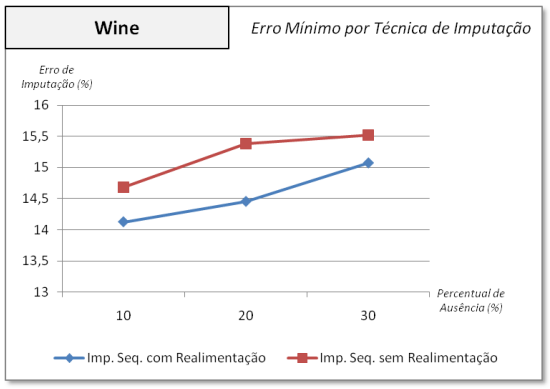


FIG. 4.9: Wine Dataset - erro de imputação

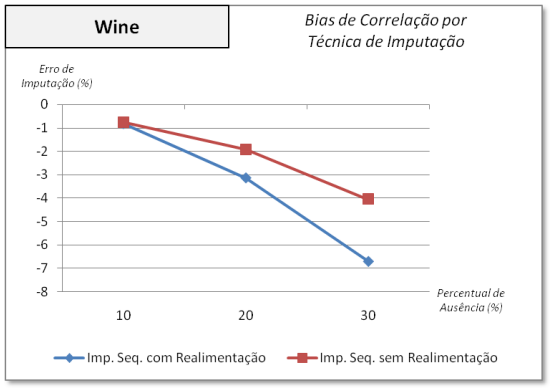


FIG. 4.10: Wine Dataset - desvio de correlação

4.2.6 DISCUSSÃO DOS RESULTADOS

A discussão dos resultados relaciona os dados apresentados pelos experimentos às questões levantadas na introdução. Para facilitar a leitura, as mesmas são reapresentadas:

- Se o reuso de valores é capaz de beneficiar o processo de imputação, aumentando o grau de precisão dos valores aferidos para substituição;
- Se o reuso de valores realmente promove um aumento artificial entre a correlação dos atributos da base imputada;
- Se o reuso de valores consegue atenuar os conhecidos efeitos negativos de uma distribuição muito ampla dos valores ausentes;
- Se a ordem de imputação dos atributos pode influenciar significativemente o resultado final do processo.

Sobre a precisão da imputação e o reuso de valores, é possível concluir que para o conjunto de bases selecionadas, a técnica de reutilização aumentou o grau de precisão do processo, uma vez que apresentou as menores taxas de erro na maioria dos casos. Esta melhora foi observada em maior grau nas bases que apresentam maior grau de correlação entre os atributos, como a Iris Dataset e a Breast Cancer Dataset. O mesmo efeito foi relatado em imputação de bases genômicas por Kim (KIM, 2004).

Em relação ao impacto sobre a correlação dos atributos, é possível constatar que independente da reutilização de valores, existe um desvio dos valores originais. Nesta avaliação, dentro do conjunto de bases selecionadas, não é possível concluir se a reutilização de valores realmente promove um fortalecimento artificial na correlação entre os atributos, como indicado por Schafer (SCHAFER, 1998). O que se pode observar nos resultados é que sempre que o desvio de correlação é positivo (i.e. a correlação é fortalecida), os processos de imputação com e sem reuso possuem desempenho similar, com o reuso de dados oferecendo índices ligeiramente melhores. Já quando o resultado é negativo (i.e. a correlação é enfraquecida), o reuso de valores apresenta um desvio significativamente maior.

Sobre os efeitos negativos de uma distribuição muito ampla de valores ausentes, é possível observar que nas bases em que o reuso de valores foi capaz de contribuir com o processo de imputação, esta melhora foi mais significativa com os percentuais de ausência mais altos (20% e 30%), ou seja, quanto mais valores ausentes, mais vantajoso é o reuso. Como é justamente nestes percentuais que aparece a dificuldade da construção dos conjuntos de treino (GELMAN, 2007), entende-se que a imputação com reuso tenha encontrado mais oportunidades de contribuição, enriquecendo continuamente os conjuntos de treino a cada iteração do processo de imputação. Isto pode ser observado na imputação da base Wine, em 20% de valores ausentes, e nas Pima e Machine, em 30%.

Tal fato não é uma surpresa total. Devido à natureza multivariada da ausência, os valores ausentes podem ocorrer nos próprios casos doadores. Quando isso acontece, os casos podem ser descartados, ou podem ter sua similaridade calculada utilizando apenas os atributos disponíveis. Se os casos doadores possuem valores ausentes em atributos com alto grau de correlação, pode ser difícil distinguir entre casos realmente similares, e casos que parecem similares, já que a falta de alguns atributos pode distorcer o cálculo de similaridade. Este tipo de cálculo pode levar a imputação de casos inconsistentes ou impossíveis, como "homens grávidos" (VANBUUREN, 2006).

Ao reutilizar os valores imputados no processo imputação sequencial, é possível preencher gradualmente os valores ausentes que ocorrem nos conjuntos de treino, aumentando as chances de se encontrar casos doadores melhores, e com maior similaridade ao caso a ser imputado.

Finalmente, em relação às cinco ordens experimentadas para imputação sequencial, não foi possível identificar diferenças significativas entre quaisquer delas, mesmo na ordenação aleatória. Como a maioria dos trabalhos relacionados que foram levantados não varia a ordem de imputação, fixando-a a partir do atributo com menor quantidade de valores ausentes (LEPKOWSKI, 2001; OUDSHOORN, 1999; KIM, 2004; VERBOVEN, 2007; IRWIN, 1994), não foi possível contextualizar estes experimentos.

Uma exceção é a tese de doutorado de Ferlin (FERLIN, 2008), que propõe a Imputação em Cascata. Neste trabalho a imputação é feita em dois níveis, sendo um primeiro baseado no agrupamento e ordenação de registros, para um posterior ordenamento dos atributos. A autora demonstra que a ordem de primeiro nível, baseada nos grupos de registros, possui um impacto significativo no processo de imputação, e que, neste nível, a utilização de ordens aleatórias é desvantajosa.

5 TRABALHOS RELACIONADOS

5.1 WEKA

O Weka (WITTEN, 2005) é sem dúvida uma das mais conhecidas plataformas de KDD. Possui diversos componentes relacionados as diferentes etapas de KDD. Os seus componentes de pré-processamento de dados são denominados filtros, e apesar de fornecer vários deles, não existem algoritmos para lidar com imputação multivariada de valores ausentes.

O Weka fornece quatro diferentes mecanismos de interface:

- **Linha de Comando**: Os componentes podem ser invocados por linha de comando, o que possibilita chamadas externas, remotas, ou utilização por scripts de batch;
- **Explorador**: Nesta interface o usuário pode navegar por diferentes menus, onde pode escolher entre os componentes disponíveis, estes menus são divididos por um esquema conceitual dos métodos de KDD, como classificação, agrupamento, associação, etc;
- **Experimentador**: Este módulo oferece ao usuário a possibilidade de experimentar variações de um mesmo componente, como a troca de parâmetros e de dados de entrada, ao final, executa análises comparativas entre as experiências. Só funciona com um componente de cada vez;
- **Knowledge Flow**: Permite a montagem e execução visual de workflows com alguns dos componentes oferecidos, inclusive com geração de gráficos de resultados. Trabalha diretamente com uma instância de workflow;

A figura 5.1 ilustra o desenho de workflows dentro da ferramenta:

5.2 TANAGARA

O Tanagara (RAKOTOMALALA, 2005) é uma ferramenta que permite a construção de processos de KDD através do encadeamento de componentes em uma árvore de execução, de maneira similar a um workflow. A figura 5.2 ilustra o funcionamento:

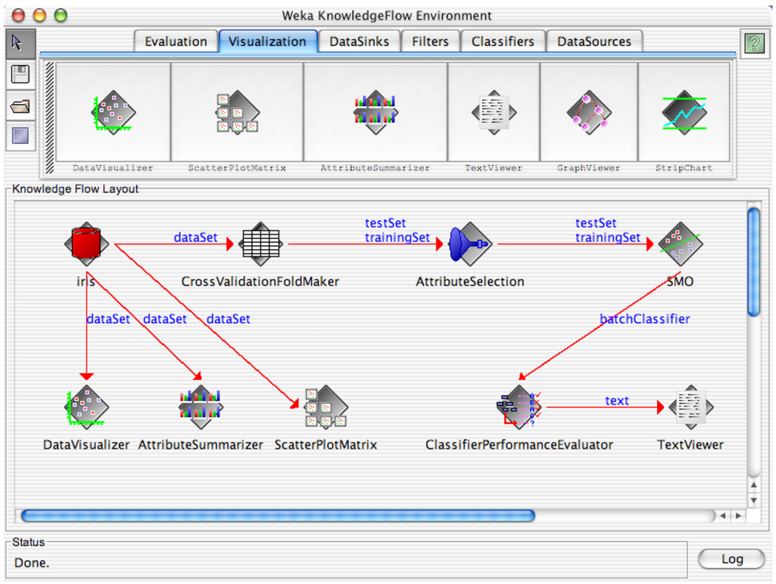


FIG. 5.1: Criação de Workflows no Weka

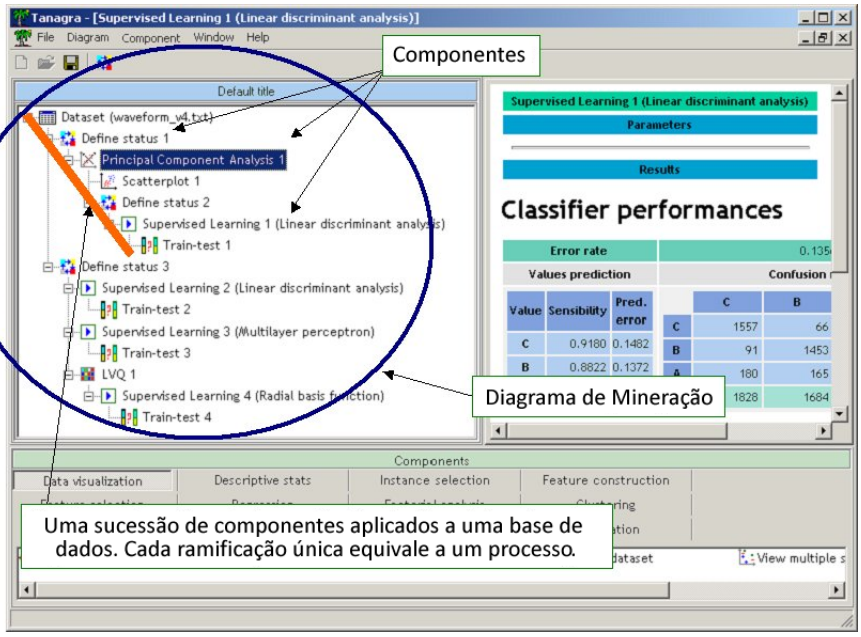


FIG. 5.2: Encadeamento de Componentes no Tanagara

Como pode ser visto, dentro de uma ramifiação da árvore é possível reaproveitar definições anteriores, e criar variações de experimento com facilidade. Em outras ferramentas que trabalham diretamente com instâncias, haveria um maior esforço, manual, para configuração de cada processo único.

Entre seus componentes, o tanagara conta com algoritmos para diversas tarefas, tais como clusterização, classificação e associação de dados. Em pré-processamento, oferece rotinas estatísticas capazes de imputar valores, como a regressão linear (JOHNSTON, 1972). Nenhuma das implementações apresentadas, porém, lida com ausência multivariada de dados.

5.3 KEPLER

O projeto Kepler (ALTINTAS, 2004) é uma plataforma para workflows científicos, que tem como objetivo auxiliar cientistas na construção e execução de wokflows, utilizando tecnologias emergentes em computação de grade. É um dos mais famosos projetos em sua área, e recebe a colaboração de diversos pesquisadores e instituições acadêmicas.A figura 5.3 ilustra a edição visual de workflows no Kepler.

A figura apresenta a execução de um workflow para retirar medidas estatísticas simples, de um vetor numérico. O primeiro componente é uma constante de dados, definida pelo usuário, e neste caso, é onde são inseridos os valores do vetor. A saída do primeiro componente é encadeada ao segundo, um extrator de medidas estatísticas simples, como a média. Das cinco saídas do segundo componente, três delas (a média, a variância, e o desvio padrão), são encadeadas a componentes de texto, que disparam as janelas observadas à direita da tela.

Apesar de não possuir um foco específico em mineração de dados, é possível encontrar adaptações feitas por pesquisadores independentes, para integração com os componentes do Weka (REUTEMANN, 2005).

5.4 VISTRAILS

O VisTrails é um sistema para construção e execução de workflow científicos, e gerência de proveniência de sistemas, em desenvolvimento pela Universidade de Utah (BAVOIL, 2005). Seu ponto forte é um sofisticado sistema de visualização de dados, especialmente preparado para atender as necessidades da biomedicina, que permite aos usuários um

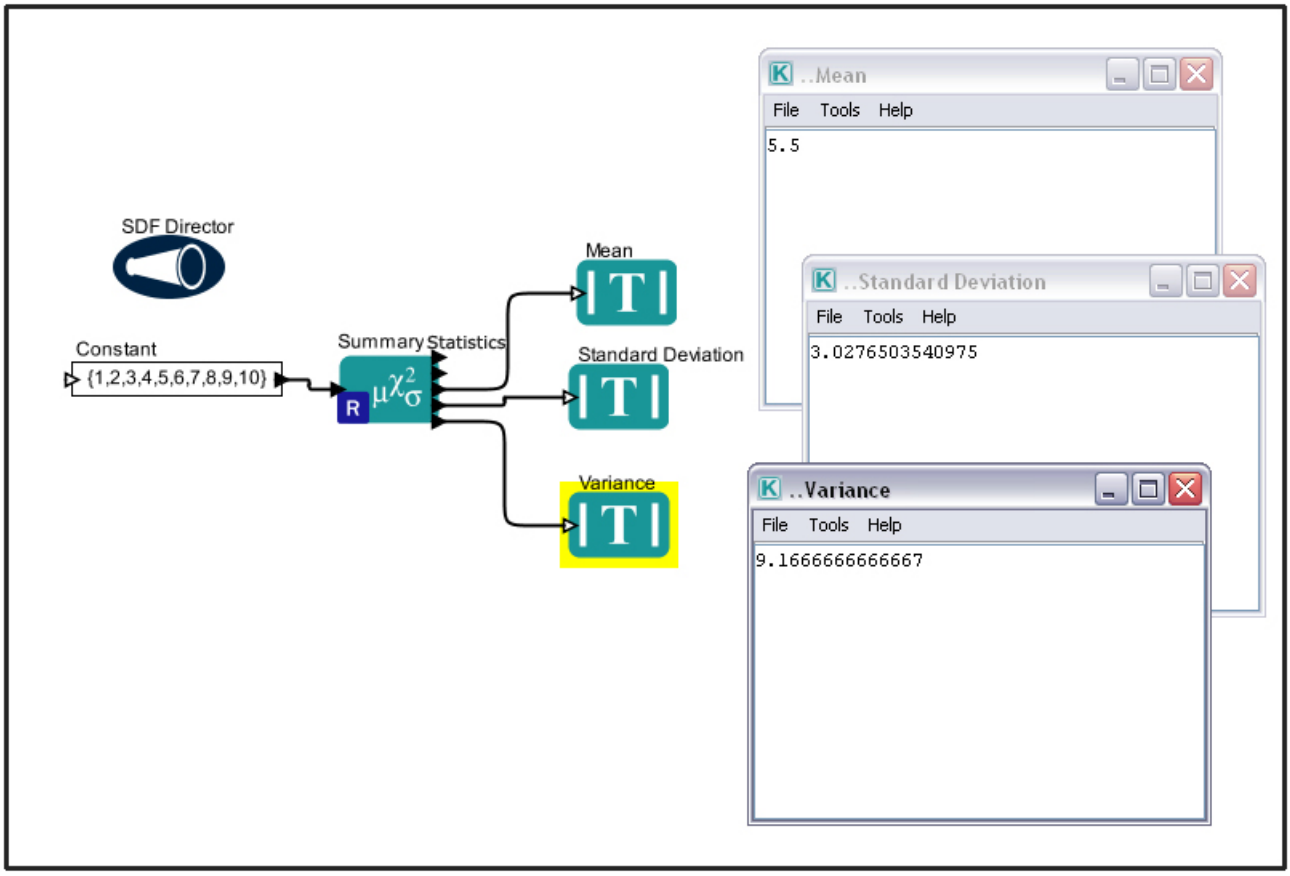


FIG. 5.3: Workflow Simples Desenhado no Weka (KEPLER, 2005)

melhor entendimento e visualização da informação. A figura 5.4 ilustra a utilização do software.

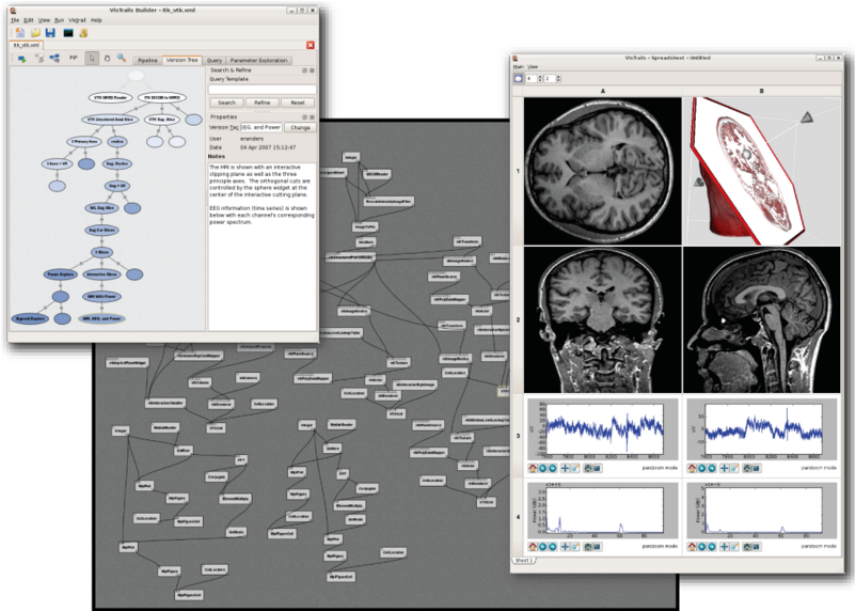


FIG. 5.4: Utilização do Vistrails em Estudos de Memória Humana (VISTRAILS, 2005)

Na ilustração é apresentado um workflow para mensuramento de dados em estudos de memória. O workflow combina etapas de processo em diversas bibliotecas diferentes, e no final é capaz de apresentar uma visualização amigável dos dados envolvidos.

Diferente dos outros sistemas workflows, e de maneira similar ao ambiente proposto neste trabalho, o Vistrails oferece um recurso para abstração e construção automática de instâncias de workflow, denominado "*pipeline*". Um *pipeline* é similar a uma definição de workflow (WFMC, 1999), no sentido de que representa apenas a estrutura do workflow, mas não define os parâmetros de execução de seus componentes. As instâncias de *pipeline* são o resultado da variação destas configurações de parâmetros, e são automaticamente construídas e executadas pela ferramenta.

5.5 SRMI

O algoritmo Sequential Regression Multivariate Imputation (SRMI) foi uma das primeiras propostas a aplicar um mecanismo efetivo de reuso de valores imputados (LEPKOWSKI, 2001). Ele funciona construindo modelos preditivos para cada coluna da base de dados, começando da coluna com menos valores ausentes para a coluna com mais valores ausentes. A cada iteração, os valores descobertos pelo modelo preditivo anterior são aproveitados na

geração do próximo modelo preditivo. O SRMI trabalha com o conceito de "restrictions and bounds", informações de conhecimento do analista de dados sobre os valores limítrofes que podem ser encontrados nas colunas, e regras auxiliares que limitam a ação dos modelos preditivos, p.ex.: "se idade < 18 então anos_fumante $= 0$".

Segundo os autores, muito embora a imputação seja possível sem a adição destas regras, o desempenho apresentado nestes casos é sempre inferior. Esta diferença deve ser encarada com cautela, pois ao mesmo tempo em que a introdução de conhecimento prévio do analista pode ajudar na imputação, esta "necessidade" torna o algoritmo semi-automático e frágil, a partir do momento em que o analista forneça informações errôneas, ou simplesmente não conheça a fundo o domínio da base de dados.

A ferramenta IVEWare (RAGHUNATHAN, 2007) é a implementação mais conhecida deste algoritmo, disponível para download gratuitamente. Desenvolvida em C e Fortran, ela funciona através de linhas de comando e arquivos de texto, que servem tanto para fornecimento dos dados de entrada e configuração do algoritmo, como para geração dos dados e avisos de saída.

A configuração é realizada através de uma linguagem própria de script, que o analista deve se familiarizar antes de trabalhar com a ferramenta. Alternativamente, o programa pode ser acoplado a plataforma de imputação SAS (SAS, 2008) como um plugin.

5.6 MICE

Outro trabalho de imputação multivariada é o algoritmo Multivariate Imputation by Chained Equations (MICE) (OUDSHOORN, 1999). O MICE trabalha com a construção de equações encadeadas, onde a imputação de um campo é considerada uma variável cuja resolução é dada por uma fórmula. Quando os valores ausentes se apresentam de maneira multivariada, as equações para cada variável são encadeadas, de maneira que a resolução de um atributo alimenta a equação e, por conseguinte, a resolução de outro. Segundo os próprios autores, dada a natureza extremamente encadeada da resolução do problema, o algoritmo não apresenta bom desempenho quando os valores ausentes ocorrem em muitas colunas, fato que prejudica a construção e o encadeamento das equações.

O MICE só existe para download como plugin do STAT (SAS, 2008), e é configurado através dos meios próprios do ambiente.

5.7 CONSIDERAÇÕES SOBRE OS TRABALHOS RELACIONADOS

A principal diferença para a implementação deste trabalho frente a outras encontradas na literatura, reside no fato de que a plataforma apresentada não é preparada apenas para realização direta de tarefas de imputação, mas também para automatizar a configuração, a execução e análise comparativa de diversos experimentos e variações de experimentos.

A tabela 5.5 relaciona o ambiente desenvolvido com os trabalhos relacionados.

Em relação ao Weka, destaca-se que embora ele possua, de maneira similar a abordagem proposta, mecanismos tanto para experimentação como para construção de workflows, os dois não funcionam em conjunto, ou seja, não é possível experimentar combinações de dados e parâmetros de maneira encadeada, mas apenas em componentes isolados.

Já sobre o Tanagara, pode-se dizer que é uma ferramenta flexível o suficiente para que se adicionem diferentes variações de um mesmo componente na árvore de execução, como por exemplo, uma entrada para o algoritmo dos K-Vizinhos com K igual 1, e outra com K igual a 3.

Tais variações porém, são sempre limitadas a um determinado ramo da árvore de configurações, e não são completamente combinadas com outras variações de parâmetros e componentes que estejam em ramificações independentes, ou em outras palavras, ramificações irmãs. O resultado final é menos eficiente do que realiza o ambiente apresentado neste trabalho, onde as variações de parâmetros são realizadas através de combinações automáticas entre todos os parâmetros indicados para experimentação.

Considerando o Kepler e o VisTrails, é importante ressaltar que são sistemas orientados para o desenvolvimento de workflows científicos, e atendem a diversos requisitos especiais, tais como (LUDÄSCHER, 2006):

- **Acesso transparante a recursos e serviços** - Esse é um requisito muito comum, normalmente atendido pela construção de serviços web capazes de fornecer um mecanismo uniforme e remoto de acesso às chamadas para o sistema de workflow.

- **Reuso de Workflows** - Determina a possibilidade da reutilização dos componentes desenvolvidos em diferentes fluxos, bem como a composição de novos workflows a partir de workflows menores.

- **Escalabilidade** - Suporte a processamento em larga escala, através de grades ou

	Weka	*Tanagara*	*SRMI (IVEWare)*	*MICE*	*Kepler*	*Vistrails*	*Ambiente Proposto*
Distribuição do Software	*Aplicativo*	*Aplicativo*	*Plugin SAS/STAT Aplicativo*	*Plugin SAS/STAT*	*Aplicativo*	*Aplicativo*	*Aplicativo*
Linguagem de Programação	*Java*	*Delphi*	*C / FORTRAN*	*C*	*Java*	*Pvthon*	*Java*
Construção de Workflows	*Apenas uma Instância por Vez*	*Encadeamento de Componentes*	*Não*	*Não*	*Apenas uma Instância por Vez*	*Meta-Construção de Instâncias*	*Meta-Construção de Definições e Instâncias*
Manipulação Visual de Workflows	*Construção e Visualização*	*Construção e Visualização*	*Não*	*Não*	*Construção e Visualização*	*Construção e Visualização*	*Apenas Visualização*
Recursos de Cache no Workflow	*Não*	*Não*	*Não*	*Não*	*Não*	*Sim*	*Cache Automática em Memória ou Disco*
Imputação Multivariada	*Não*	*Não*	*Sim*	*Sim*	*Não*	*Não*	*Sim*
Reuso de Dados Imputados	*Não*	*Não*	*Sim - Obrigatório*	*Não - Obrigatório*	*Não*	*Não*	*Opcional, com qualquer Algoritmo de Imputação*
Fácil Criação de Componentes	*Sim*	*Sim*	*Não*	*Não*	*Não*	*Não*	*Sim*
Fontes de Dados Suportadas	*ARFF*	*CSV*	*CSV*	*CSV*	*ODBC, JDBC, CSV*	*ODBC, CSV*	*ODBC, JDBC, Excel, CSV e ARFF*
Visualização Gráfica de Dados	*Sim*	*Sim*	*Não*	*Não*	*Sim*	*Sim*	*Não*
Abrangência dentro da área de KDD	*Pré-Processamento, Mineração de Dados e Pós-Processamento*	*Pré-Processamento, Mineração de Dados e Pós-Processamento*	*Pré-Processamento*	*Pré-Processamento*	*Não é específico para KDD*	*Não é específico para KDD*	*Pré-Processamento*

FIG. 5.5: Comparação entre os trabalhos relacionados

clusters computacionais.

- **Execução Independente** - Os workflows de longa duração devem oferecer um modo de execução no qual funcionam em máquinas remotas, de maneira independente da máquina cliente que os iniciou, ou das que os monitoram.

Estes e outros requisitos não são considerados para o ambiente proposto, uma vez que foram levantados para o controle de tarefas genéricas em processos científicos, que normalmente envolvem diferentes formatos de dados, diversos pesquisadores envolvidos, e grande quantidade de informação (LUDÄSCHER, 2006), enquanto a elaboração deste trabalho atende a uma processo pontual, e muito mais específico.

Sobre a criação automática de instâncias, muito embora presente no VisTrails, ela é oferecida com menos flexibilidade do que a apresentada neste trabalho, que consiste na combinação total dos parâmetros envolvidos na definição de workflow. Outra diferença é que o ambiente proposto é capaz de gerar automaticamente instâncias e definições, trabalhando em um nível superior de abstração, mais próximo ao processo de negócio da imputação, enquanto o VisTrails requer a construção manual das definições, para somente depois criar as instâncias.

Sobre o SRMI e o MICE, fica claro que são inciativas voltadas para a execução de um único algoritmo, e não possuem aspirações maiores do que se comportarem como um componente de imputação. Neste sentido, é possível que venham a ser adicionadas ao ambiente no futuro.

As principais limitações da abordagem proposta estão na interface gráfica, que só permite a visualização dos workflows, quando o ideal seria permitir uma construção interativa, com recursos como arrastar-e-soltar de componentes. Por enquanto os usuários ainda devem recorrer a programação por API, ou a declaração de arquivos de propriedades.

A falta de mecanismos para visualização de dados também é um ponto negativo, mas de menor importância, dada a grande quantidade existente de programas gratuitos que já realizam este tipo de trabalho. O desgaste de utilizar programas de terceiros é o constante trabalho de adaptação dos dados para os diferentes formatos exigidos por cada ferramenta.

6 CONCLUSÕES

Uma dificuldade comum aos processos de análise de dados é a existência de valores ausentes, em especial quando estes ocorrem de maneira multivariada. Procedimentos de imputação sequencial são capazes de lidar com os valores ausentes em diversas colunas, imputando valores em uma coluna por vez, porém oferecem diversas possibilidades de implementação e experimentação, entre as quais analistas de dados muitas vezes têm dificuldade em escolher.

Com tanta diversidade, é possível encontrar divergências e questões em aberto na literatura, relativas a realização de imputação sequencial, como por exemplo, sobre o reuso de valores no processo de imputação. Um dos principais fatores deste problema é a falta de metodologias e ferramentas próprias para o controle, experimentação e avaliação dos métodos de imputação.

Buscando solucionar estas questões, este trabalho apresentou uma abordagem de combinação prática e teórica dentro do campo da Imputação Sequencial em cenários multivariados, na forma de uma metodologia para imputação, e sua implementação como um ambiente baseados nos principais conceitos da teoria de workflows. O ambiente proposto foi utilizado para realização de testes em imputação sequencial, a fim de avaliar questões relacionadas ao reuso de valores na imputação.

Com a utilização do ambiente proposto, foi possível automatizar diversos experimentos que de outra maneira seriam manualmente executados, um-a-um. Ao estabelecer definições de workflow comuns para os testes com e sem reuso, foi ainda possível garantir que a única diferença era realmente o fator de reutilização, isolando os experimentos de outros fatores que pudessem distorcer a avaliação final.

6.1 LISTA DE CONTRIBUIÇÕES

- **A definição de uma metodologia para execução e avaliação de técnicas de imputação, capaz de apoiar novos trabalhos de experimentação**: Com o estabelecimento da metodologia foi possível compreender e explorar separadamente três fases da experimentação, a geração de valores ausentes, a imputação dos valores, e a análise dos resultados. No caso deste trabalho foi explorada a imputação

sequencial, porém a metodologia não se restringe ao mesmo, e pode ser modificada para outras abordagens de imputação pré-existentes, ou na experimentação de novas (FERLIN, 2008).

- **Desenvolvimento de um ambiente que implemente a metodologia proposta, utilizando conceitos de workflows:** O ambiente de imputação foi capaz de atender os requisitos levantados, configurando, executando e avaliando 1800 experimentos diferentes com apenas cinco configurações. O custo de se replicar todos os experimentos manualmente, utilizando mecanismos comuns de outras ferramentas conhecidas, seria extremamente oneroso para o analista. O benefício da utilização de uma estrutura de workflows capaz de automatizar os testes em todos os aspectos (configuração, execução, e análise), é cada vez maior na medida em que a complexidade dos experimentos cresce. Ainda, ao trabalhar com uma estrutura de componentes e fluxos, modificar um experimento se tornou uma tarefa simples e passível de automatização. Criar novas variantes de métodos já incorporados na plataforma se mostrou uma operação mais segura, pois os componentes reaproveitados já haviam sido testados e utilizados em produção. Na experimentação, foram empregados dois métodos principais de imputação, um com reuso de valores, e outro sem. Uma vez que praticamente toda a estrutura entre os dois métodos era comum, o ambiente garantiu que a única diferença entre o método proposto e o método original era o fator de reuso de dados, concretizado como uma etapa opcional no workflow. Isto permite a exclusão, nos experimentos, de fatores comuns de incerteza de comparação, tais como qualidade da implementação de diferentes algoritmos, diferenças entre linguagens e plataformas de programação e carregamento dos dados, entre outros, que tradicionalmente dificultam a avaliação entre diferentes técnicas de imputação.

- **Apresentação de resultados experimentais que permitam concluir, no domínio de um conjunto determinado de bases de dados, se o reuso de valores pode aprimorar a qualidade final dos dados imputados; se o reuso de valores promove um aumento articifial da correlação entre atributos; se a ordem de imputação impacta de maneira significativa o processo de imputação:** Os resultados mostram que o reuso de valores pode promover uma melhoria significativa nos casos onde existe correlação na base a ser imputada.

Destacaram-se, para as bases experimentadas, as seguintes vantagens no uso de um mecanismo de reutilização de valores:

i Preservar a consistência de imputação entre as diferentes colunas, sem abrir mão dos benefícios advindos da decomposição de um problema multivariado em vários problemas univariados;

ii Diminuir progressivamente a ocorrência geral de valores ausentes, uma vez que em cada iteração novos valores são preenchidos;

iii Minimizar o risco da imputação de casos incoerentes;

iv Facilitar, a cada iteração, o trabalho de algoritmos como K-Vizinhos, Back Propagation e Redes Bayiseanas, que são extremamente prejudicados pela ocorrência de valores ausentes em casos preditivos.

Sobre a correlação dos atributos, é possível constatar que qualquer processo de imputação exerce um desvio dos valores originais. Segundo Schafer (SCHAFER, 1998), o reuso de dados provoca um aumento artifical dos fatores de correlação, porém o que se observa nos resultados é que os processos de imputação com e sem reuso possuem um comportamento similar, sempre que o desvio de correlação é positivo, com vantagem para o reuso de valores. Quando o resultado é negativo porém, o reuso de valores apresenta um desvio significativamente maior. Por último, dentro do contexto dos experimentos não foi possível tecer conclusões acerca da ordenação dos atributos, e do seu impacto no processo de imputação, uma vez que os resultados obtidos pelas diferentes ordens foi muito próximo.

6.2 TRABALHOS FUTUROS

Algumas questões referentes ao trabalho apresentado nesta disseração ainda podem ser mais desenvolvidos. Assim, as seguintes orientações para a continuidade deste trabalho podem ser propostas:

a) O desenvolvimento de uma interface gráfica, é possível tornar a utilização do ambiente mais amigável, sem a necessidade de se trabalhar com programação, ou de se estudar o formato de arquivos de configuração pré-determinados. Tal funcionalidade pode facilitar a adoção do ambiente por outros pesquisadores, interessados em colaborar no desenvolvimento de novos componentes.

b) O suporte a processamento paralelo/distribuído é vital para a imputação em grandes bases de dados, ou para a condução de muitos experimentos. Em Ferlin (FERLIN, 2008), onde o ambiente foi utilizada, realizaram-se 252.000 experimentos diferentes, em 15 computadores. Muito embora a grande quantidade de experimentos só tenha se tornado possível pela utilização do ambiente apresentado nesta dissertação, a manipulação de 15 computadores diferentes se mostrou uma tarefa desgastante, que poderia ser automatizada com o suporte a processamento paralelo ou distribuído.

c) A integração com componentes de outras ferramentas, apesar de já acontecer no ambiente, ainda pode melhorar significativamente, dada a quantidade de componentes pré-implementados em diversos trabalhos relacionados. O Weka, por exemplo, pois possui diversos filtros de pré-processamento e componentes de visualização de dados, que poderiam enriquecer a experiência do analista na construção dos workflows de imputação, e na análise dos resultados dos experimentos, dentro do ambiente proposto.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALTINTAS, I., BERKLEY, C., JAEGER, E., JONES, M., LUDÄSCHER, B. e MOCK, S. **Kepler: An Extensible System for Design and Execution of Scientific Workflows**. Em *In SSDBM*, págs. 21–23, 2004.

BAVOIL, L., CALLAHAN, S. P., CROSSNO, P. J., FREIRE, J. e VO, H. T. **VisTrails: Enabling interactive multiple-view visualizations**. Em *In Proc. IEEE Visualization 2005*, págs. 135–142, 2005.

BAYENS, T. **State of Workflow**, 2004. URL `http://www.theserverside.com/tt/articles/article.tss?l=Workflow`. The Server Side.COM.

CASTANEDA, R., FERLIN, C., GOLDSCHMIDT, R., SOARES, J., CARVALHO, L. e CHOREN, R. **Aprimorando Processos de Imputação Multivariada de Dados com Workflows**. Em *XXIII Simpósio Brasileiro de Banco de Dados - São Paulo - Brasil*. 2008.

CINGOLANI, P. **JSwarm-PSO: Particle Swarm Optimization**, 2005. URL `http://sourceforge.net/projects/jswarm-pso/`.

COHEN, M. P. e HUANG, G. G. **Analyzing Survey Data by Complete-Case and Available-Case Methods**. págs. 289–294, 2000.

COMISSION, U. N. S. e FOR EUROPE, E. C. **Glossary of Terms on Statistical Data Editing**, 2000. United Nations.

COMPANY, H. M. **Webster's New International Dictionary of the English Language - Second Edition**. Houghton Mifflin Company, 2004.

FARHANGFAR, A., KURGAN, L. e PEDRYCZ, W. **A Novel Framework for Imputation of Missing Values in Databases**. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(5):692–709, 2007.

FARIS, P., GHALI, W., BRANT, R., NORRIS, C., GALBRAITH, P. e KNUDTSON, M. **Multiple Imputation Versus Data Enhancement for Dealing with Missing Data in Observational Health Care Outcome Analyses**. *Journal of Clinical Epidemiology*, 55:184–191, 2002.

FERLIN, C. **Imputação em Cascata: Uma abordagem para imputação multivariada de dados**. Tese de Doutorado, COPPE/UFRJ, 2008.

GAMMA, E., HELM, R., JOHNSON, R. e VLISSIDES, J. **Design patterns: Elements of reusable object-oriented software**. Addison Wesley, 1995.

GELMAN, A. e HILL, J. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. Cambridge University Press, 2006.

GELMAN, A., LEVY, M. e ABAYOMI, K. **Diagnostics for Multivariate Imputations**, 2007. Social Science Research Network - Social Science Electronic Publishing, Inc.

GELMAN, A. e RAGHUNATHAN, T. **Conditionally Specified Distributions: An Introduction**. *Statistical Science*, 16(3):268–269, 2001.

GLEASON, T. e STAELIN, R. **A Proposal for Handling Missing Data**. *Psychometrika*, 40(2):229–252, 1974.

GOLDSCHMIDT, R. e PASSOS, E. **Data Mining: Um Guia Prático - Conceitos, Técnicas, Ferramentas, Orientações e Aplicações**, volume 1. Editora Campus, Rio de Janeiro, 1 edition, 2005.

HEERINGA, S., LITTLE, R. e RAGHUNATHAN, T. **Multivariate Imputation of Coarsened Survey Data on Household Wealth**. Em *Survey Nonresponse*, págs. 357–371. 2002.

HOLLINGSWORTH, D. **The Workflow Reference Model: 10 Years On**, 2004. URL `http://www.wfmc.org/standards/docs/Ref_Model_10_years_on_Hollingsworth.pdf`. Fujitsu Services, UK; Technical Committee Chair of The Workflow Management Coalition Specification.

IRWIN, M., COC, N. e KONG, A. **Sequential Imputation or Multilocus Linkage Analisys**. *Proceedings of the National Academy of Sciences of the United States of America*, 91(24):11684–11688, 1994. ISSN 0027-8424.

JOHNSTON, J. **Econometric methods**. McGraw-Hill, 1972.

JONSSON, P. e WOHLIN, C. **An Evaluation of k-Nearest Neighbour Imputation Using Likert Data**. Em *10th International Symposium on Metrics*, págs. 108–118. 2004.

KENNICKELL, A. **Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances**, 1997. The Survey of Consumer Finances Bibliography, The Federal Reserve Board.

KEPLER. **Kepler Getting Started Guide**, 2005. URL `http://www.kepler-project.org/Wiki.jsp?page=Documentation`.

KIM, K.-Y., KIM, B.-J. e YI, G.-S. **Reuse of imputed data in microarray analysis increases imputation efficiency**. *BMC Bioinformatics*, 5(1):160, 2004. ISSN 1471-2105. URL `http://www.biomedcentral.com/1471-2105/5/160`.

LAKSHMINARAYAN, K., HARP, S. e SAMAD, T. **Imputation of Missing Data in Industrial Databases**. *Applied Intelligence*, 11(3):259–275, 1999.

LEPKOWSKI, J., RAGHUNATHAN, T., SOLENBERGER, P. e VAN HOEWYK, J. **A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models**. *Statistics Canada*, 27(1):85–95, 2001.

LITTLE, R. e RUBIN, D. **Statistical Analysis With Missing Data**. *Technometrics*, 45:364–365, 2003.

LUDÄSCHER, B., ALTINTAS, I., BERKLEY, C., HIGGINS, D., JAEGER, E., JONES, M., LEE, E. A., TAO, J. e ZHAO, Y. **Scientific workflow management and the Kepler system: Research Articles**. *Concurrency Computation Pract. Exper.*, 18 (10):1039–1065, 2006. ISSN 1532-0626.

MAGNANI, M. **Techniques for Dealing with Missing Data in Knowledge Discovery Tasks**, 2004. University of Bologna, Department of Computer Science.

MARRONE, P. **Joone: Java Object Oriented Neural Engine - The Complete Guide**, 2007. URL `http://www.joone.org`.

MERZ, C. e MURPHY, P. **UCI Repository of Machine Learning Databases**, 1998. University of California, Irvine, Department of Information and Computer Sciences. http://www.ics.uci.edu/ mlearn/MLRepository.html.

MONARD, M. C. **An analysis of four missing data treatment methods for supervised learning**. *Applied Artificial Intelligence*, 17:519–533, 2003.

N., R. e MEFFERT, K. **JGAP: The Java Genetic Algorithms Package**, 2007. URL `http://jgap.sourceforge.net`.

NSDL.ORG. **The National Science Digital Library**, 2008. http://nsdl.org/.

OUDSHOORN, C., VAN BUUREN, S. e VAN RIJCKEVORSEL, J. **Flexible multiple imputation by chained equations**, 1999. Netherlands Organization for Applied Scientific, Technical Report PG/VGZ/99.045.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. e HOEWYK, J. V. **IVEware - Imputation and Variance Estimation Software**, 2007. URL `http://www.isr.umich.edu/src/smp/ive`. Survey Methodology Program, Institute for Social Research - University of Michigan.

RAKOTOMALALA, R. **TANAGRA : un logiciel gratuit pour l'enseignement et la recherche**. *Actes de EGC*, 2:697–702, 2005.

REUTEMANN, P. **Kepler and Ptolemy II - KeplerWeka**, 2005. URL `http://www.cs.waikato.ac.nz/ fracpete/projects/kepler_and_ptolemy`.

ROYSTON, P. **Multiple Imputation of Missing Values: Update of Ice**. *Stata Journal*, 5(2):188–201, 2005.

RUBIN, D. **Nested Multiple Imputation of NMES via Partially Incompatible MCMC**. *Statistica Neerlandica*, 57(1):3–18, 2003.

SAS. **Statistical Analysis with SAS/STAT® Software**, 2008. URL http://www.sas.com/technologies/analytics/statistics/stat/index.html.

SCHAFER, J. **Analisys of Incomplete Multivariate Data**, volume 1. Chapman and Hall-CRC, Rio de Janeiro, 1 edition, 1997.

SCHAFER, J. L. e OLSEN, M. K. **Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective**. *Multivariate Behavioral Research*, 33(4):545–71, 1998. ISSN 0027-3171.

SCHÖNER, H. **Working with Real-World Datasets**. Technische Universität Berlin, 2004.

SLOMINSKI, A. **Introduction to Workflows and Use of Workflows in Grids and Grid Portals**, 2003. Indiana University.

SOARES, J. **Pré-Processamento em Mineração De Dados: Um Estudo Comparativo em Complementção**. Tese de Doutorado, COPPE/UFRJ, 2007.

TEKNOMO, K. **K-Nearest Neighbors Tutorial**, 2004. URL http://people.revoledu.com/kardi/tutorial/KNN.

VAN BUUREN, S., BRAND, J., GROOTHUIS-OUDSHOORN, C. e RUBIN, D. **Fully Conditional Specification in Multivariate Imputation**. *Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

VERBOVEN, S., BRANDEN, K. V. e GOOS, P. **Sequential imputation for missing values**. *Comput. Biol. Chem.*, 31(5-6):320–327, 2007. ISSN 1476-9271.

VISTRAILS. **Vistrails Documentation**, 2005. URL http://www.vistrails.org/index.php/Documentation.

WFMC. **Workflow Management Coalition Terminology and Glossary**, 1999. URL http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf. The Workflow Management Coalition Specification.

WITTEN, I. H. e FRANK, E. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2005.

YI, X., ALLAN, J. e LAVRENKO, V. **Discovering Missing Values in Semi-Structured Databases**. Em *8th RIAO Conference Proceedings*. 2007.

YUAN, J. **Multiple Imputation for Missing Data: Concepts and New Development**, 2008. SAS Institute Technical Report P267-25.