UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

IMPUTAÇÃO DE DADOS BASEADO EM OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS CONSIDERANDO OS PRINCIPAIS MECANISMOS DE AUSÊNCIA DE DADOS

LILIAN DE JESUS CHAVES DIAS

DM:15/2013

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2013

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LILIAN DE JESUS CHAVES DIAS

IMPUTAÇÃO DE DADOS BASEADO EM OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS CONSIDERANDO OS PRINCIPAIS MECANISMOS DE AUSÊNCIA DE DADOS

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2013

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LILIAN DE JESUS CHAVES DIAS

IMPUTAÇÃO DE DADOS BASEADO EM OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS CONSIDERANDO OS PRINCIPAIS MECANISMOS DE AUSÊNCIA DE DADOS

Dissertação submetida à Banca Examinadora do Programa de Pós-graduação em Engenharia Elétrica da UFPA para o obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada, elaborada sob a orientação do Prof. Dr. Ádamo Lima de Santana.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2013

Dados Internacionais de Catalogação-na-Publicação (CIP)

Dias, Lilian de Jesus Chaves, 1989-
Imputação de dados baseado em otimização por enxame de partículas considerando os principais mecanismos de ausência de dados / Lilian de Jesus Chaves Dias. - 2013.

Orientador: Ádamo Lima de Santana.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2013.

1. Inteligência artificial. 2. Inteligência coletiva. 3. Otimização matemática. 4. Simulação. I. Título.

CDD 22. ed. 006.3

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

IMPUTAÇÃO DE DADOS BASEADO EM OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS CONSIDERANDO OS PRINCIPAIS MECANISMOS DE AUSÊNCIA DE DADOS

AUTOR: LILIAN DE JESUS CHAVES DIAS

DISSERTAÇÃO DE MESTRADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADDA ADEQUADA PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.
APROVADA EM 18/06/2013

BANCA EXAMINADORA:

Prof. Dr. Ádamo Lima de Santana
(ORIENTADOR – UFPA)

Profa. Dra. Adriana Rosa Garcez Castro
(MEMBRO – UFPA)

Prof. Dr. Cláudio Alex Jorge da Rocha
(MEMBROI FPa)

Prof. Dr. **Prof. Dr. Nandamudi Lankalapalli Vijaykumar**
(MEMBRO INPE)

VISTO:

Prof. Dr. Evaldo Gonçalves Pelaes
(COORDENADOR DO PPGEE/ITEC/UFPA)

AGRADECIMENTOS

Gostaria de agradecer, antes de todos, a minha mãe, Inalda Maria Chaves Dias, por todas as conversas, discursões, abraços, beijos e amor durante todos os dias da minha vida.

Também agradeço ao meu orientador, o Prof. Dr. Ádamo Lima de Santana, por ter me dado esta oportunidade de aprimorar meus conhecimentos acadêmicos, pelo suporte, dedicação e companheirismo durante esses anos de dedicação ao Mestrado.

Agradeço aos integrantes do Laboratório de Inteligência Computacional e Pesquisa Operacional – LINC da UFPa, pelo apoio, paciência, conversas e companheirismo, em especial ao Fábio, Arilene, Vincent, Nathália e Ivan, agradeço imensamente por ter convivido com vocês.

Agradeço também aos integrantes do Laboratório de Planejamento de Redes e de Alto Desempenho – LPRAD da UFPa pelo convívio, companheirismo, amizade e conselhos durantes os anos de Mestrado, minha vida se tornou mais prazerosa com a presença de vocês nela.

Agradeço ao Prof. Dr. Otávio Noura Teixeira do Cesupa, por ter instigado a ideia de realizar o Mestrado, por ter me introduzido o caminho dos Algoritmos Genéticos e da linha acadêmica, o senhor tem a minha eterna admiração.

Gostaria de agradecer aos meus familiares e amigos pela paciência e pela sua existência, já que sem vocês a vida seria sem gargalhadas e momentos inesquecíveis.

SUMÁRIO

LISTA DE FIGURAS VIII

LISTA DE TABELAS IX

LISTA DE SIGLAS X

RESUMO XI

ABSTRACT XII

1 INTRODUÇÃO 1

1.1 MOTIVAÇÃO 1

1.2 OBJETIVOS E METODOLOGIA 2

1.3 ESTRUTURA DO TRABALHO 3

2 TRATAMENTO DE VALORES AUSENTES 5

2.1 CONSIDERAÇÕES INICIAIS 5

2.2 DEFINIÇÕES 5

2.3 MECANISMOS DE AUSÊNCIA DE DADOS 7

2.4 TRATAMENTO DE VALORES AUSENTES 8

2.4.1 IMPUTAÇÃO SIMPLES 9

2.4.2 IMPUTAÇÃO MÚLTIPLA 10

2.5 SÍNTESE DO CAPÍTULO 11

3 OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS 12

3.1 CONSIDERAÇÕES INICIAIS 12

3.2 COMPUTAÇÃO NATURAL 12

3.3 OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS 14

3.4 SÍNTESE DO CAPÍTULO 17

4 TRABALHOS CORRELATOS 19

4.1 CONSIDERAÇÕES INICIAIS 19

4.2 ESPECIFICAÇÃO DO MECANISMO DE VALORES AUSENTES NOS EXPERIMENTOS 19

4.3 MÉTODOS DE IMPUTAÇÃO BIOINSPIRADOS 20

4.4 SÍNTESE DO CAPÍTULO 24

5 TRATAMENTO DE VALORES AUSENTES UTILIZANDO ENXAME DE PARTÍCULAS 25

5.1 CONSIDERAÇÕES INICIAIS 25
5.2 TRATAMENTO DE VALORES AUSENTES UTILIZANDO ENXAME DE PARTÍCULAS 25
5.2.1 MODELAGEM DA PARTÍCULA 26
5.2.2 FUNÇÃO DE APITIDÃO 28
5.2.3 CÁLCULO DA VELOCIDADE E INÉRCIA 29
5.3 SÍNTESE DO CAPÍTULO 30
6 CONFIGURAÇÕES E REALIZAÇÃO DOS EXPERIMENTOS 31
6.1 CONSIDERAÇÕES INICIAIS 31
6.2 CONFIGURAÇÕES DOS EXPERIMENTOS 31
6.3 CONFIGURAÇÕES DOS MÉTODOS DE IMPUTAÇÃO UTILIZADOS 34
6.4 FUNÇÃO DE AVALIAÇÃO 35
6.5 SÍNTESE DO CAPÍTULO 36
7 RESULTADOS DOS EXPERIMENTOS 37
7.1 CONSIDERAÇÕES INICIAIS 37
7.2 RESULTADOS E ANÁLISE DOS EXPERIMENTOS 37
7.3 SÍNTESE DO CAPÍTULO 44
8 CONCLUSÃO 45
REFERÊNCIAS 48
APÊNDICE A – LISTA DOS RESULTADOS DOS EXPERIMENTOS POR BASE 53
APÊNDICE B – CARACTERÍSTICAS DAS BASES DE DADOS COM VALORES AUSENTES CONSIDERANDO OS MECANISMOS DE AUSÊNCIA 61

LISTA DE FIGURAS

FIGURA 1. FLUXOGRAMA DE UM PSO. ----------16
FIGURA 2. PROCEDIMENTO DE CRIAÇÃO DA LISTA DE VALORES DE DOMÍNIO E DA PARTICULAR DO ALGORITMO. ----------27
FIGURA 3. GRÁFICO DA BASE VERTEBRAL-COLUMN.----------38
FIGURA 4. GRÁFICO DOS RESULTADOS DO MECANISMO MCAR PARA A BASE TIC-TAC-TOE. --------39
FIGURA 5. GRÁFICO DOS RESULTADOS DO MECANISMO MAR PARA A BASE CONTRACEPTIVE. ----39
FIGURA 6. GRÁFICO DOS RESULTADOS DO MECANISMO NMAR PARA A BASE ÍRIS. ----------40
FIGURA 7. GRÁFICO DOS MELHORES RESULTADOS OBTIDOS PELOS MÉTODOS DE TRATAMENTO DE VA CONSIDERANDO DIFERENTES PORCENTAGENS DE AUSÊNCIA NO ATRIBUTO. ----------41
FIGURA 8. GRÁFICO DOS MELHORES RESULTADOS OBTIDOS PELOS MÉTODOS DE TRATAMENTO DE VA CONSIDERANDO OS MECANISMOS DE AUSÊNCIA DE DADOS. ----------42
FIGURA 9. GRÁFICO DOS MELHORES RESULTADOS OBTIDOS PELOS MÉTODOS DE TRATAMENTO DE VA EM GERAL. ----------43

LISTA DE TABELAS

TABELA 6.1. PROPRIEDADES DAS BASES DE DADOS SELECIONADAS PARA A REALIZAÇÃO DOS EXPERIMENTOS. 31
TABELA 6.2. CARACTERÍSTICAS DAS BASES SINTÉTICAS GERADAS. 32
TABELA 6.3. CONFIGURAÇÃO DO TRATAMENTO DE VALORES AUSENTES POR ENXAME DE PARTÍCULAS. 34
TABELA 6.4. CONFIGURAÇÃO DO CLASSIFICADOR 3-NN. 34
TABELA 6.5. CONFIGURAÇÃO DO CLASSIFICADOR C4.5. 34
TABELA 6.6. CONFIGURAÇÕES DOS MÉTODOS DE IMPUTAÇÃO KNNIMPUTE E SVMIMPUTE. 35

LISTA DE SIGLAS

TVA – Tratamento de Valores Ausentes

VA – Valores Ausentes

MCAR – *Missing Completely At Random*

MAR – *Missing At Random*

NMAR – *Not Missing At Random*

KNN – *K-Nearest Neighbor*

SVM – *Support Vector Machine*

KNNImpute – Imputação de dados por KNN

SVMImpute – Imputação de dados por SVM

KEEL – *Knowledge Extraction based on Evolutionary Learning*

PSO – *Particle Swarm Optimization*

RNA – Redes Neurais Artificiais

AG – Algoritmo Genético

MLP – *Multilayer Perceptron*

WEKA – *Waikato Environment for Knowledge Analysis*

RMSE – *Root Mean Square Error*

RESUMO

Durante o processo de extração do conhecimento em bases de dados, alguns problemas podem ser encontrados como por exemplo, a ausência de determinada instância de um atributo. A ocorrência de tal problemática pode causar efeitos danosos nos resultados finais do processo, pois afeta diretamente a qualidade dos dados a ser submetido a um algoritmo de aprendizado de máquina. Na literatura, diversas propostas são apresentadas a fim de contornar tal dano, dentre eles está a de imputação de dados, a qual estima um valor plausível para substituir o ausente. Seguindo essa área de solução para o problema de valores ausentes, diversos trabalhos foram analisados e algumas observações foram realizadas como, a pouca utilização de bases sintéticas que simulem os principais mecanismos de ausência de dados e uma recente tendência a utilização de algoritmos bioinspirados como tratamento do problema. Com base nesse cenário, esta dissertação apresenta um método de imputação de dados baseado em otimização por enxame de partículas, pouco explorado na área, e o aplica para o tratamento de bases sinteticamente geradas, as quais consideram os principais mecanismos de ausência de dados, MAR, MCAR e NMAR. Os resultados obtidos ao comprar diferentes configurações do método à outros dois conhecidos na área (KNNImpute e SVMImpute) são promissores para sua utilização na área de tratamento de valores ausentes uma vez que alcançou os melhores valores na maioria dos experimentos realizados.

Palavras-chave: *Tratamento de valores ausentes, mecanismo de ausência de dados, imputação de dados, valores ausentes, valores faltosos, PSO, enxame de partículas.*

ABSTRACT

During the knowledge discovery in database process some problems may be found, e.g. some instance of one attribute may be missing. Such issue can even cause harmful effects to the final results of the process, since directly affects the data quality of a database which some machine learning algorithm may be applied to. In the literature are some proposals to solve such harm; among them is the data imputation process that estimates a plausible value to fill in the missing one. Inside the area of missing value treatment, some researches were analyzed and observations were raised such as, a few utilization of synthetic datasets that simulates the main mechanisms of missingness and a tendency to use bioinspired algorithm to treat the missing values. From this scenario, the present dissertation analyses an imputation method based on particle swarm optimization, an underexplored one, and applies it to the treatment of synthetics datasets generated considering the main mechanisms of missingness, MAR, MCAR and NMAR. The results obtained when comparing the algorithm against different configurations of itself and another two treatments known in the area (KNNImpute and SVMImpute) are promising for its use as missing value treatment whereas the bioinspired method reached the bests values for the major of the experiments.

Keywords: *Missing value treatment, mechanism of missingness, data imputation, missing data, PSO, particle swarm optimization.*

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Em um conjunto de dados, um pesquisador pode se deparar com o problema de ausência de valores para determinados instâncias de alguns atributos (FACELI et al., 2011). Essa ausência pode ter diversas causas como por exemplo: incorreta inserção manual de dados, medidas incorretas, erro em equipamentos, dentre outros. E em algumas áreas, chega ser comum encontrar bases com valores ausentes em mais de 50% de suas entradas (FARHANGFAR;KURGAN; PEDRYCZ, 2007) (LAKSHMINARAYAN, HARP e SAMAD, 1999). A falta de alguns exemplos pode ocasionar problemas durante a análise de dados uma vez que esta precisa da maior quantidade disponível de exemplos para descobrir a existência de um determinado padrão na base de dados. Exemplos de problemas ocasionados pela ausência de dados podem ser, a perda de eficiência do método para análise de dados e a imposição de viés na base (FARHANGFAR, KURGAN e PEDRYCZ, 2004).

A fim de solucionar tal problemática, algumas alternativas têm sido propostas como utilização da média, utilização para análise apenas de casos completos, eliminação dos objetos com valores ausentes e estimação de valores para substituir os ausentes - chamado de imputação de dados (SCHAFER e GRAHAM, 2002) (BROWN e KROS, 2003).

Quando focado neste último tipo de solução, uma análise de publicações foi realizada com aplicações em áreas diversas, e verificou-se uma falta de padronização quanto a origem da base de dados a ser utilizada nos experimentos realizados para os tratamentos de valores ausentes (DIAS, LOBATO e SANTANA, DE, 2013). Sobre tal observação, as bases comumente utilizadas são oriundas de repositórios públicos como o UCI (ASUNCION e NEWMAN, 2007), porém elas podem ou não possuir valores ausentes em seus dados. Usualmente escolhem-se bases completas para depois remover, controladamente, valores dessa base a fim de analisar o comportamento de um determinado método, esse processo origina as chamadas bases sintéticas. Com isso, os pesquisadores conseguem simular alguns dos mecanismos de ausência existentes na literatura da área. Os mais replicados são o *Missing At Random* (MAR) – o qual a ausência de valores

depende de um atributo presente na base; e *Missing Completely At Random* (MCAR) – onde a ausência de um exemplo é completamente independente dos valores observados. Porém, para melhorar o desenvolvimento da área é preciso realizar análises além desses dois mecanismos usuais, tornando-se um desafio a análise do tratamento para os três mecanismos existentes: MAR, MCAR e *Not Missing At Random* (NMAR) - a ausência de valores pode depender de um valor observável e/ou ausente na base. Nesse cenário, são raros os trabalhos que conseguem, efetivamente, realizar um estudo desse tipo, principalmente quando deseja-se uma análise mais extensa, utilizando uma determinada quantidade de bases por exemplo.

Além da observação descrita acima, outra tendência se destaca na área de tratamento de valores ausentes (TVA), a qual consiste na utilização de métodos bioinspirados seja como auxiliadores de outros métodos ou como tratamento propriamente dito. Sendo assim alguns recentes trabalhos mostram promissores resultados para esse tipo de tratamento, porém ainda muito incipiente se comparado com os métodos mais usuais da área (AZADEH et al., 2013; FIGUEROA GARCÍA, KALENATIC e LOPEZ BELLO, 2011; FRANÇA, DE, COELHO e ZUBEN, VON, 2013; VERONEZE et al., 2011).

Com base nesses dois pontos, esta dissertação se propõe a desenvolver um sistema de imputação de dados, onde o algoritmo bioinspirado de otimização por enxame de partículas é adaptado para o tratamento de valores ausentes. Além disso, também é realizada uma análise entre cinco tratamentos diferentes – dois usuais da área (SVMImpute e KNNImpute) e os outros três sendo compostos de diferentes configurações do algoritmo bioinspirado; a fim de verificar seu comportamento em sete bases de dados distintas, as quais foram sinteticamente geradas de acordo com os mecanismos de ausência de dados variando sua porcentagem de atributos ausentes. Os objetivos e a metodologia desta dissertação são apresentados na seção a seguir.

1.2 OBJETIVOS E METODOLOGIA

O principal objetivo da dissertação é o de analisar o comportamento de um algoritmo de otimização por enxame de partículas adaptado para o problema de valores ausentes utilizando

diversas bases de dados e os diferentes mecanismos de ausência de dados; e como consequência, os seguintes objetivos específicos:

- Modelar o algoritmo de otimização por enxame de partículas para o tratamento de valores ausentes;
- Comparar o método proposto com outros já conhecidos na literatura da área.

Para cumprir com tais objetivos, a seguinte metodologia foi utilizada:

- Estudo das definições de dados ausentes;
- Estudo dos mecanismos de ausência dos dados;
- Estudo dos principais métodos de tratamento;
- Estudo teórico sobre Computação Natural;
- Estudo sobre otimização por enxame de partículas;
- Modelagem da partícula de acordo com o problema de valores ausentes;
- Codificação do algoritmo de otimização por enxame de partículas;
- Estudo de métricas de comparação usuais da área;
- Execução dos experimentos;
- Consolidação dos resultados; e
- Análise dos resultados.

A seguir, uma estrutura é apresentada com o intuito de guiar o leitor pelo assuntos necessários para melhor entender esta dissertação.

1.3 ESTRUTURA DO TRABALHO

Os primeiros capítulos desta dissertação realizam a fundamentação teórica dos assuntos mais utilizados. No Capítulo 2, os principais conceitos relacionados a dados ausentes são explicados assim como os métodos tratamento que são usados durante a análise do algoritmo apresentado por esta dissertação. O Capítulo 3 apresenta os pontos mais relevantes sobre Computação Natural a fim de esclarecer sobre o algoritmo bioinspirado escolhido para esse

trabalho, o de otimização por enxame de partículas. No Capítulo 4 são mostrados os trabalhos correlatos a este, sendo explicitamente apontados as diferenças entre as propostas. O Capítulo 5 apresenta como o tratamento de valores ausentes utilizando o algoritmo de otimização por enxame de partículas foi modelado e codificado. Já Capítulo 6 as configurações dos experimentos são apresentadas, assim como a origem das bases utilizadas. O Capítulo 7 apresenta os resultados dos experimentos realizados seguido de uma análise. E por fim, no Capítulo 8 são feitas as considerações finais e são apresentadas algumas perspectivas para trabalhos futuros nessa linha de pesquisa.

No próximo capítulo, uma fundamentação teoria mais detalhada sobre tratamento de valores ausentes é apresentada.

2 TRATAMENTO DE VALORES AUSENTES

2.1 CONSIDERAÇÕES INICIAIS

Neste capítulo são descritas as principais definições de dados ausentes, assim como os mecanismos de ausência encontrados quando uma base apresenta tal problema. Também são descritos as principais categorias para seu tratamento, bem como os métodos de imputação que são utilizados nesta dissertação.

2.2 DEFINIÇÕES

Um objeto em um conjunto de dados é dito completo se todos os seus atributos estão preenchidos de forma apropriada. Um dado ausente, dado faltoso, dado incompleto ou valor ausente (VA), indica que um exemplo de um objeto não está preenchido. Todas essas nomenclaturas são provenientes dos termos em inglês, *Missing Data* ou *Incomplete Data*, e se referem a mesma definição (LITTLE e RUBIN, 2002)(SILVA, 2010)(VERONEZE, 2011)(FACELI et al., 2011)(GONÇALVES DE OLIVEIRA, 2009). Ou ainda como Mcknight et al. (2007) menciona, o termo dados faltantes, significa que está faltando algum tipo de informação sobre o fenômeno que está sendo analisado.

Em estatística, o principal objetivo das análises na área é realizar inferências válidas sobre uma população de interesse, com ou sem a ocorrência de valores ausentes. Todavia, suas principais técnicas não toleram esse tipo de problema. Com isso, os valores ausentes tem como consequência a imposição de viés ao processo, e isso ocorre tanto em análises que, de alguma forma, consideram os dados ausentes, quanto quando são tratados (SCHAFER e GRAHAM, 2002). Vale ressaltar que tal problema também está presente no processo de análise de dados (HEERINGA, WEST e BERGLUND, 2010).

Sobre as causas que levam à ausência de dados, são diversas as razões para tal, sendo principalmente dependentes do mecanismo de aquisição de dados. Brown & Kros (2003) apresentam alguns exemplos que podem ocasionar tal problema:

- Fatores operacionais: erros na entrada dos dados, estimativas, remoção acidental de campos de tabelas, entre outras;
- Recusa na resposta em pesquisas;
- Impossibilidade de aplicação de um determinado questionamento.

Após a apresentação desses conceitos, esta dissertação adotou a fundamentação apresentada em Little & Rubin (2002) como base teórica. Sendo assim, a ocorrência de valores ausentes em uma variável está relacionada a algum processo identificável, ou seja, é possível conhecer os mecanismos causadores de tal ausência. Os quais após descobertos podem auxiliar na escolha da técnica de tratamento mais adequada.

Com isso em mente, Rubin (1976) apresenta algumas definições para melhor entender a causa da ausência de dados no processo de análise de dados. Seja $U = (U_1,\ldots, U_n)$ um vetor de variável aleatória com função densidade de probabilidade f_θ. O objetivo é realizar inferências sobre θ, o vetor parâmetro dessa densidade. Seja $M = (M_1,\ldots, M_n)$ um vetor de variável aleatória associado ao "indicadores de dados ausentes", onde cada M recebe o valor 0 ou 1. A probabilidade de M tomar o valor $m = (m_1, \ldots, m_n)$ dado que U toma o valor $u = (u_1, \ldots, u_n)$ é $g_\phi(m|u)$, onde ϕ é o parâmetro vetor de estorvo da distribuição.

A distribuição condicional g_ϕ corresponde ao "processo que causa a ausência dos dados": se $m_i = 1$, o valor da variável aleatória U_i será observado, enquanto $m_i = 0$, o valor de U_i não será observado (RUBIN, 1976). Em Mcknight *et al. (*2007), é utilizado uma nomenclatura diferente para o vetor M, o qual é denominado *Dummy codes*, possuindo mesma dimensão dos dados observados em U, onde $m_i = 0$ caso u_1 seja observado, e $m_1 = 1$, caso contrário. Ainda é possível desmembrar o vetor U em dois conjuntos $U = \{U_{obs}, U_{aus}\}$, onde U_{obs} são os dados observados e U_{aus} os ausentes. Essa última é a nomenclatura adota por este trabalho.

As definições acima descritas, fornecem o insumo para entender a relação entre as causas dos dados faltoso, denominado de Mecanismos de ausência de dados, do inglês, *Mechanism of Missingness*, os quais são mais detalhados a seguir.

2.3 MECANISMOS DE AUSÊNCIA DE DADOS

A forma apropriada de se tratar os valores ausentes está interligada em como os exemplos foram removidos. O mecanismo de ausência de dados realiza o mapeamento dessas condições estatisticamente, e é caracterizado pela distribuição condicional de **M** dado **U**:

$$\boldsymbol{p}(\mathbf{M}|\mathbf{U},\boldsymbol{\xi}) = \boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}},\mathbf{U}_{\boldsymbol{aus}},\boldsymbol{\xi}) \quad (1)$$

onde ξ , denota o parâmetro desconhecido que define um dos três mecanismos de ausência de dados propostos por Little & Rubin (2002). São eles:

- *Missing completely at random* (MCAR)*:* situação que ocorre quando a probabilidade da variável ser ausente é independente da própria variável ou de qualquer outra influência. A condição para ser MCAR é expressa pela relação abaixo:

$$\boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}},\mathbf{U}_{\boldsymbol{aus}},\boldsymbol{\xi}) = \boldsymbol{p}(\mathbf{M}|\boldsymbol{\xi}) \quad (2)$$

o que mostra que a ausência da variável não depende dos valores de entrada pois, os exemplos disponíveis contém toda a informação para realizar inferências sobre os dados ausentes. Logo, a razão para a ausência de dados é completamente aleatória – a probabilidade de uma observação ser ausente não é relacionada a qualquer outra característica encontrada nas instâncias.

- *Missing at random* (MAR): a ausência de dados é independente dos valores ausentes, mas o padrão de ausência é predita por outras variáveis observáveis da base de dados. A condição para ser considerada MAR é expressa pela relação:

$$\boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}}, \mathbf{U}_{\boldsymbol{aus}}, \boldsymbol{\xi}) = \boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}}, \boldsymbol{\xi}) \quad (3)$$

onde a ausência da variável depende apenas de valores observados nos dados de entrada (casos completos).

- *Not Missing at random* (NMAR)*:* o padrão de dados faltosos não é aleatório e pode depender tanto do próprio valor ausente, quanto de um valor presente na base. Little & Rubin (2002) descreve esta ideia por meio da equação:

$$\boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}}, \mathbf{U}_{\boldsymbol{aus}}, \boldsymbol{\xi}) \neq \boldsymbol{p}(\mathbf{M}|\mathbf{U}_{\boldsymbol{obs}}, \boldsymbol{\xi}) \quad (4)$$

em contraste com o padrão MAR, a variável ausente no caso NMAR não pode ser predita apenas levando-se em consideração as variáveis do conjunto de dados. O que torna este o mecanismo mais difícil de ser reconhecido e predito.

A respeito dos mecanismos MCAR e MAR, é comum denomina-los de padrões *ignoráveis*. Tal nomenclatura é devido a facilidade desses mecanismos serem manipulados, uma vez que sua estimação pode ser feita considerando os valores disponíveis para análise (SCHAFER, 1997) (MCKNIGHT et al., 2007) (GRAHAM, 2009).

Em compensação, o padrão NMAR é chamado de *não-ignorável*, uma vez que não há informação dentro do conjunto de dados observáveis que auxilia na estimação de seus valores. Por conta disso, a construção de um modelo, estatístico ou de aprendizado de máquina, para realizar tal estimação torna-se complicada.

Na literatura, a maioria das pesquisas envolvendo essa problemática, assumem que os valores ausentes estão no padrão MAR ou MCAR.

2.4 TRATAMENTO DE VALORES AUSENTES

Baseado no problema de valores ausentes previamente descrito, pesquisadores se propuseram, ao longo dos anos, a solucionar tal problema utilizando-se de métodos estatísticos ou

não para isso. Little & Rubin (2002) categorizaram os métodos usualmente utilizados na área da seguinte maneira:

- Ignorando e descartando registros e atributos incompletos;
- Estimação de parâmetros na ocorrência de valores ausentes;
- Procedimentos de imputação de dados.

Entretanto, autores como Brown & Kros (2003) citam uma categoria a mais, chamada de deleção de casos ou atributos selecionados; enquanto que em Farhangfar, Kurgan e Pedrycz (2007) são apontados apenas duas dessas categorias: 1) remoção dos dados ausentes e 2) imputação dos dados ausentes, caracterizando uma falta de consenso sobre tais categorias. Este trabalho utiliza as definições conforme Little & Rubin (2002). Mais precisamente, apenas o conceito de procedimentos de imputação de dados tem alguns de seus métodos mais detalhados nessa seção.

Os procedimentos de imputação de dados são baseados no princípio de preenchimento de valores ausentes por outros dados disponíveis (BROWN e KROS, 2003). As técnicas de imputação de dados aqui apresentados estão divididos em dois macro grupos: imputação simples e imputação múltipla, podendo haver diversas técnicas para cada um.

2.4.1 IMPUTAÇÃO SIMPLES

Nesse tipo de técnica, seu princípio básico é imputar um valor único para cada dado ausente na base de dados, analisando-o posteriormente, como se não houvesse dados ausentes, gerando uma base imputada (MCKNIGHT et al., 2007). São exemplos desse tipo de técnica a Imputação por K-vizinho mais próximo[1] e por máquinas de vetor de suporte[2].

[1] do inglês K-Nearest Neighbor (KNN)
[2] do inglês Support Vector Machines (SVM)

- Imputação por K-vizinho mais próximo - *KNNImpute* (BATISTA e MONARD, 2003). Esse método é baseado em instância, assim, toda vez que um valor ausente é encontrado em uma linha da base, o algoritmo calcula o *k* vizinho mais próximo, e um valor a partir desses é imputado. Para valores nominais, o valor mais comum entre todos os vizinhos é usado, e para valores numéricos, o valor médio. Portanto, é necessária uma medida de proximidade entre as instâncias para que possa ser definido. A distância Euclidiana é a mais utilizada.
- Imputação por máquinas de vetor de suporte - *SVMImpute* (HONGHAI, GUOSHUN e CHENG, 2005). Esse método é baseado em regressão. Primeiramente o algoritmo seleciona exemplos de atributos que possuem um valor que não esteja ausente. Depois, o método define um dos atributos de condição (atributo de entrada), que possua valores ausentes, como sendo o atributo de decisão (atributo de saída), e os atributos de decisão como os atributos de condição, para, em seguida, usar a regressão para prever os valores do atributo de decisão.

2.4.2 IMPUTAÇÃO MÚLTIPLA

Os algoritmos desse grupo funcionam não somente para restaurar a variabilidade natural nos dados ausentes, mas também para incorporar a incerteza causada pela estimação dos valores ausentes (WAYMAN, 2003).

Neste procedimento, os valores ausentes para uma variável de um determinado objeto são estimados usando os valores observados de outros objetos para o mesmo atributo. Esses valores são imputados, gerando uma base completa sem valores ausentes, sendo este processo realizado múltiplas vezes, gerando múltiplas bases imputadas. Em cada uma das bases, uma análise estatística é realizada produzindo resultados de análises múltiplas. Tais análises são combinadas e geram uma análise geral (WAYMAN, 2003).

Baseado nos dois grupos de imputação dos dados, este trabalho utiliza as definições de imputação simples, por ser a mais utilizada na área e possuir diversos métodos já bem difundidos na literatura. Especificamente, são utilizados os algoritmos descritos na seção 2.4.1, os quais

foram usados durante os experimentos desta dissertação por meio da utilização da ferramenta KEEL (ALCALÁ-FDEZ et al., 2008), e tiveram seus resultados comparados com os obtidos pelo método de tratamento aqui apresentados e o qual será melhor explicado no Capítulo 5.

2.5 SÍNTESE DO CAPÍTULO

Neste capítulo foram apresentados as principais definições do problema de tratamento de valores ausentes, encontrado em diversas áreas de aplicação. Também foram mostrados os mecanismos de ausência que uma base pode assumir, MAR, MCAR e NMAR, e os quais são simulados neste trabalho.

Além disso, foram categorizados os possíveis tratamentos para o problema, dando maior enfoque a imputação simples, sendo os algoritmos de imputação por K-vizinho mais próximo e máquinas de vetor de suporte descritos. E para finalizar, uma breve introdução à imputação múltipla.

Ressalta-se que os algoritmos descritos neste capítulo são os utilizados nesta dissertação para análise de seu comportamento quando submetidos a imputação em bases de dados com valores ausentes. As quais, são sinteticamente geradas para simular mecanismos de ausência brevemente explicados.

A seguir é apresentada a fundamentação teórica de Otimização por Enxame de Partículas, afim de elucidar os principais pontos que compõe o algoritmo, o qual foi adaptado para o problema de valores ausentes.

3 OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS

3.1 CONSIDERAÇÕES INICIAIS

Este capítulo apresenta uma breve introdução sobre computação natural e suas ramificações. Principalmente sobre a chamada computação bioinspirada, e duas de suas abordagens, a computação evolucionária e a inteligência de enxames. Sendo esta última mais detalhadamente descrita para melhor entender um de seus algoritmos mais conhecidos, o de otimização por enxame partículas, do inglês *particle swarm optimization* (PSO), o qual é adaptado nessa dissertação para o problema de valores ausentes.

3.2 COMPUTAÇÃO NATURAL

Computadores mostraram-se úteis em diversos domínios e aplicações, como o armazenamento de informações de uma empresa e o processo de decisão que ocorre durante o pouso de um avião. Apesar desses processos não aparentarem ser tão transparentes para seus usuários finais, a capacidade de realizar essas tarefas é resultado de décadas de pesquisas. Durante décadas, também testemunhou-se o surgimento e o teste de diferentes paradigmas computacionais e a maior utilização de computadores. Desde meados da década de 1940, trabalhos com programação linear, principalmente, se beneficiaram com o desenvolvimento dos computadores (CASTRO, DE, 2006).

Outra tendência que se evidenciou em meados dos anos de 1940, e a qual recebeu mais atenção nas últimas duas ou três décadas, é a ideia de fusão da natureza e da computação. Mais especificamente, existem estudiosos que usam modelagens científicas baseadas na natureza em técnicas de resolução de problemas, para realizar a síntese de eventos biológicos (bioinformática), e até incorporando-as em materiais computacionais (nanotecnologia). Tais

fatores constituem a chamada Computação Natural, a qual possui as seguintes ramificações, consideradas as principais e sendo nomeadas conforme segue (CASTRO, DE, 2006):

- Computação bioinspirada: faz uso da natureza como forma de inspiração para o desenvolvimento de técnicas de resolução de problemas. Sua ideia principal consiste em inspira-se por meio da natureza para resolver problemas complexos a fim de desenvolver ferramentas computacionais, ou algoritmos.
- Simulação e emulação da natureza por meio da computação: essa ramificação é basicamente um processo sintético que visa criar padrões, formas, comportamentos e organismos os quais, não necessariamente, se assemelham a "vida-como-nós-conhecemos". Seus produtos podem ser usados para simular vários fenômenos naturais, aumentando assim a compreensão da natureza e as percepções sobre modelos computacionais;
- Computação com materiais naturais: corresponde ao uso de novos materiais para realizar cálculos, constituindo assim um novo paradigma de computação, que surge para substituir ou complementar os computadores atuais à base de silício.

Acerca do trabalho aqui descrito, apenas o conceito de computação bioinspirada será utilizado, tendo suas principais especificações e exemplos descritos no decorrer deste capítulo.

Dentre as abordagens da computação natural, algoritmos e sistemas computacionais inspirados na natureza são os mais antigos e popularmente utilizados exemplos da área, como em McCulloch & Pitts (1943) onde foi proposto o primeiro modelo matemático de um neurônio, originando as Redes Neurais Artificiais, ou RNA (BISHOP, 1996; HAYKIN, 1999).

Os algoritmos e sistemas baseado em computação natural possuem dois objetivos principais. O primeiro é a do interesse de pesquisadores em modelar fenômenos naturais, e fazer a sua simulação em computadores, ou seja, um modelo capaz de simular algo observado na natureza e que consiga ser reproduzido por meio de linhas de código e interpretado por uma máquina com funcionalidades semelhantes ao original. O segundo objetivo envolve o estudo do fenômeno, processo e até modelos teóricos naturais afim de conseguir entende-los e modela-los para resolução de problemas. A principal motivação para a criação desses algoritmos e sistemas computacionais, é o fato de que os métodos tradicionais não obtém uma solução satisfatória para um certo problema (CASTRO, DE, 2007).

Essas técnicas computacionais inspiradas na natureza, podem ser chamadas de computação bioinspirada ou computação motivada biologicamente (MANGE e TOMASSINI, 1998), ou ainda de computação com metáforas biológicas (PATON, 1994). Dentre as abordagens conhecidas podemos citar o algoritmo de otimização por enxame de partículas, descrito a seguir.

3.3 OTIMIZAÇÃO POR ENXAME DE PARTÍCULAS

Dentre os ramos da computação inspirada na natureza, existe o de inteligência de enxame, do inglês *Swarm Intelligence*, a qual baseia-se na observação de sociedades naturais, tanto a humana quanto a de animais e possui em sua essência a ideia de que as pessoas não aprendem somente umas com as outras, mas seus conhecimentos e habilidades se disseminam entre si, convergindo uma população para um processo ótimo. Logo, uma população pode aprender a partir de observações dos indivíduos ao seu redor (KENNEDY, EBERHART e SHI, 2001)

O objetivo dos algoritmos desse ramo é modelar o comportamento de indivíduos, a interação local desses com o ambiente que o cerca e com seus vizinhos individuais, a fim de se obter um comportamento o qual pode ser utilizado para resolver problemas mais complexos, como problemas de otimização (ENGELBRECHT, 2007).

Três são os fatores regem esses algoritmos: avaliar, comparar e imitar. O primeiro consiste nos indivíduos estimarem seu comportamento baseados na sua capacidade de sentir o ambiente ao seu redor. O segundo, os indivíduos usam uns aos outros como parâmetros de comparação. E, por fim, a imitação é importante para aquisição e manutenção das habilidades, o que converge a população para um melhor ponto (EIBEN e SMITH, 2003).

Kennedy, Eberhart e Shi (2001) mencionam que os indivíduos aprendem localmente com seus vizinhos, por meio de interações com sua vizinhança, e compartilhando experiências entre si, ocorrendo assim um processo de aprendizado. Tal fator está interligado com a disseminação do conhecimento por meio de resultados de aprendizado social. O qual pode ser observado em crenças, atitudes, comportamento e outros tipos de manifestações entre indivíduos de uma população. Com isso, percebe-se que uma sociedade é um sistema auto-organizado com

propriedades globais as quais não podem ser preditas utilizando apenas as propriedades individuais de quem a compõe (KENNEDY, EBERHART e SHI, 2001).

Por fim, outro item mencionado por eles, é que o conhecimento é otimizável por culturas. Ou seja, embora as interações sejam locais, a experiência e as inovações que ocorrem são transportadas pela cultura até os indivíduos mais afastados da população, havendo interação com esses e gerando resultados mais promissores para todos. Esse efeito global torna-se transparente a todos os indivíduos fazendo com que a sociedade utilize isso para beneficio próprio, e por consequência de toda a sua população.

Os comportamentos citados fazem parte do modelo cultural adaptativo, o qual rege a modelagem dos algoritmos de inteligência de enxames. As abordagens mais conhecidas são os trabalhos inspirados por comportamento sociais de insetos, como exemplos os algoritmos colônia de formigas, enxame de partículas, colônia de abelhas, dentre outros (KENNEDY, EBERHART e SHI, 2001).

Mais especificamente, esta dissertação fala sobre o algoritmo de otimização por enxame de partículas (PSO) pelos seguintes motivos:

- Por ser o algoritmo do ramo mais conhecido academicamente;
- Possuir diversas aplicações (POLI, 2008);
- Por conseguir se adequar a abordagem de imputação de dados;
- Poucos trabalhos dessa área de tratamento de valores ausentes o utilizam;
- Pela sua implementação ser facilmente remodelada para o processamento paralelo, compondo uma parte ainda inexplorada na área de tratamento de valores ausentes; e
- Possuir um tempo de execução otimizado se comparados com outras técnicas usadas como os Algoritmos Genéticos (AG).

A técnica do PSO é estocástica e baseada em populações, a qual foi primeiramente desenvolvida por Kennedy & Eberhart (1995). Seu algoritmo faz alusão ao comportamento social por meio da interação entre indivíduos, nesse caso partículas, de um determinado grupo, denominado de enxame.

Sua inspiração teve origem a partir da observação de bandos de pássaros e de cardumes de peixes, durante a sua busca por alimento em uma determinada região. Por meio dessa observação,

evidenciou-se que o comportamento do grupo é influenciado pela experiência individual acumulada de cada partícula, e também pela experiência acumulada do grupo como um todo em relação a sua tarefa (KENNEDY, EBERHART e SHI, 2001).

No PSO, cada possível solução do problema corresponde a um ponto no espaço de busca, ou dimensão. Essas soluções, ou partículas, por sua vez, possuem um valor associado, o qual é avaliado individualmente para cada partícula e indica o quão adequada esta se encontra para resolver o problema; além desse, cada solução possui uma velocidade que define a direção do movimento de cada partícula. Por meio da modificação da velocidade, a qual leva em consideração a posição da partícula e a melhor posição do grupo, e ao longo do tempo, o grupo consegue alcançar seu objetivo. O fluxograma da Figura 1 mostra o funcionamento de um PSO (CARACIOLO, 2009).

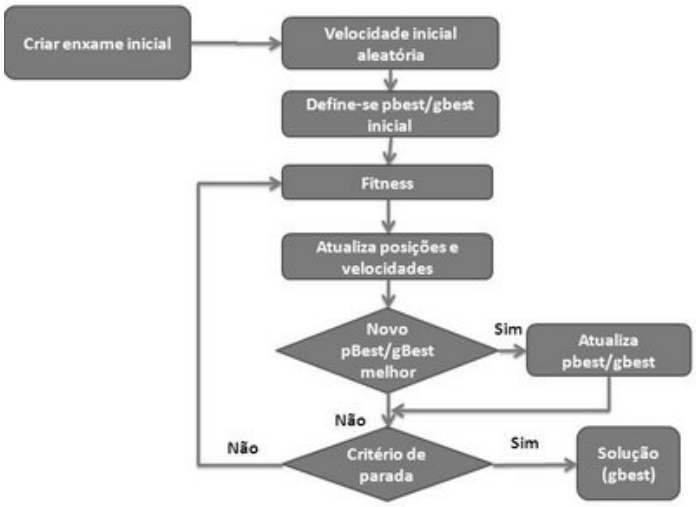


Figura 1. Fluxograma de um PSO.

Onde *gbest* e *pbest* caracterizam a melhor partícula global e local, sendo a primeira a melhor partícula do enxame e a segunda a melhor combinação já encontrada para uma determinada partícula, respectivamente.

No algoritmo, o enxame é iniciado aleatoriamente, com uma população inicial de soluções candidatas representando cada partícula, a qual tem seus valores e velocidades também iniciadas de forma aleatória. No modelo canônico do PSO a velocidade e a atualização da posição de cada partícula podem ser calculadas de acordo com a Eq.(5) e Eq. (6), respectivamente, adaptado de Engelbrecht (2007):

$$\boldsymbol{V(t+1) = V(t) + b1 * Rnd * (\mathrm{mLocal} - \mathrm{X(t)}) + b2 * Rnd * (mGlobal - X(t))} \quad (5)$$

$$\boldsymbol{X(t+1) = X(t) + V(t+1)} \quad (6)$$

Onde, na Eq. (5): *V* corresponde a velocidade da partícula na iteração t (anterior) e t+1 (atual); *Rnd* é um valor aleatório o qual varia entre [0,1]; *X(t)* é a posição anterior da partícula; *mLocal* é a melhor posição anterior da partícula; *mGlobal* é a melhor posição anterior do enxame; *b1* e *b2* são os coeficientes de atração, onde irá tender mais para o *mLocal* ou para o *mGlobal*. E a Eq.(6) demostra como a posição de cada partícula é atualizada dependendo do valor da velocidade. Logo, percebe-se que a partícula considera as melhores posições encontradas individualmente e pelo grupo para atualizar sua velocidade, e cada uma é atualizada independentemente. Vale ressaltar que a conexão entre as dimensões dá-se pela função objetivo, a qual é influenciada pelas posições das partículas. Sendo assim, quanto melhor a combinação de posição de cada dimensão, melhores são os valores encontrados pela função objetivo.

A mudança de velocidade das partículas modifica suas posições fazendo-as se movimentar através do espaço do problema. Com isso, ao longo de sucessivas iterações, os resultados da decisão individual e da influência social, exercida pelo valor do melhor do grupo, faz com que as partículas acabem convergindo para uma solução ótima (ENGELBRECHT, 2007).

A partir dessas informações percebe-se que no PSO, cada potencial solução tem a si associado uma velocidade, a qual faz com que as partículas consigam voar pelo hiperespaço do problema (EBERHART e KENNEDY, 1995). Assim, o enxame possui uma espécie de memória dos melhores resultados já visitados pela partícula e pela população inteira.

Maiores informações de como o PSO foi modelado para o tratamento de valores ausentes são descritas no Capítulo 5 desta dissertação.

3.4 SÍNTESE DO CAPÍTULO

Este capítulo apresentou um pouco sobre as principais características da área da Computação Natural. Uma delas, a Computação Bioinspirada, possui algoritmos baseados em observações provindas da natureza tenho como principal ramificação a inteligência de enxame, o qual se baseia na observação do comportamento social de indivíduos de uma população e possui exemplos como enxame de partículas e colônia de formigas.

As principais características da inteligência de enxames consiste em avaliar, comparar e imitar, o que faz com que os indivíduos da população comparem-se com seus melhores vizinhos e imitando-os conseguem alcançar o objetivo da tarefa.

Este trabalho baseia-se na implementação de um tratamento de valor ausente utilizando o PSO. A escolha deste algoritmo em detrimento de outros, ocorreu pelo fato deste possuir uma boa capacidade de adaptação ao problema, ter um menor custo computacional se comparados com outros algoritmos, por ter suas características facilmente adaptáveis ao paralelismo, campo ainda não explorado no tratamento de valores ausentes, e possuir poucos trabalhos que o utilizem como método para estimar valores ausentes.

O Capítulo a seguir abrange os trabalhos correlatos para tratamento de valores ausentes, focando-se em algumas características de seus experimentos, como quantidade de mecanismos de valores ausentes utilizados, e a utilização de algoritmos bioinspirados.

4 TRABALHOS CORRELATOS

4.1 CONSIDERAÇÕES INICIAIS

Este capítulo descreve o levantamento bibliográfico de trabalhos relevantes na literatura sobre imputação os quais consideram os diferentes mecanismos de valores ausentes, e utilizam métodos de tratamento bioinspirados, e além disso, cita as principais diferenças destes para o método utilizado nesta dissertação.

A primeira seção relata sobre trabalhos os quais especificam a utilização de bases considerando os mecanismos de valores ausentes em seus experimentos. A segunda é voltada para apontar as diferenças entre trabalhos que utilizam métodos de imputação bioinspirados.

4.2 ESPECIFICAÇÃO DO MECANISMO DE VALORES AUSENTES NOS EXPERIMENTOS

Em um trabalho de tratamento de valores ausentes, pode haver a especificação de qual mecanismo de ausência de dados a base utilizada nos experimentos ou estudo de caso se encontra. Entretanto, são poucos os que fazem isso explicitamente, havendo uma dificuldade de identificar o padrão da ausência na base. Os estudos descritos a seguir apresentam de forma explícita o tipo de mecanismo utilizado em seus experimentos, possibilitando a sua divisão em trabalhos que utilizam um ou mais de um mecanismo de valores ausentes. Vale ressaltar que são considerados apenas os experimentos realizados, caso um trabalho apresente um tratamento bioinspirado, este será citado na referente seção designada para este tipo de método.

Um exemplo de estudo que utiliza apenas um mecanismo é o de SILVA-RAMÍREZ et al., (2011). Nesse trabalho foram selecionadas quatorze bases de dados, as quais tiveram seus valores removidos de forma aleatória de acordo com uma determinada porcentagem, ou seja, a base tem *x*% de seus valores retirados de variáveis arbitrárias, caracterizando o mecanismo MCAR. As

bases geradas foram aplicadas ao método de Rede Neurais *Multilayer Perceptron* (MLP) para imputação dos valores ausentes. Sendo assim, há a comparação dos métodos utilizando apenas dos mecanismo existentes.

Outro trabalho que pode ser mencionado é o de LUENGO, SÁEZ e HERRERA, (2012), o qual, apesar de diversos algoritmos para o tratamento de valores ausentes serem usados com diferentes classificadores, as bases utilizadas eram provenientes do UCI, as quais não possuem uma especificação quanto ao mecanismo de sua ausência, por conta disso, os pesquisadores reconhecem as bases provindas desse repositório como sendo aleatórias (MAR). Logo, as análises feitas também consideram apenas um mecanismo.

Além desses, também podem ser citados trabalhos como (CHENG, LAW e SIU, 2012; DING e ROSS, 2012; LI e WANG, 2012; LIU e BROWN, 2013; ZHANG, 2012) os quais também utilizam apenas um mecanismo de valor ausente, sendo MCAR ou MAR. Nenhum dos trabalhos utiliza o mecanismo NMAR, por ser considerado o mais difícil de ter sua característica simulada nas bases e de se estimar seus valores.

Como exemplos de trabalhos que utilizam mais de um mecanismo durante seus experimentos os de CHANG, ZHANG e YAO (2012) e WOHLRAB & FÜRNKRANZ (2010) chamam atenção. O que mostra uma tendência mais atual para a análise de tratamentos de valores ausentes considerando mais de um mecanismo, porém ainda muito incipiente entre as pesquisas na área.

Com base nos trabalhos citados, percebe-se que a área de tratamento de valores ausentes possui ainda fatores pouco explorados, como o comportamento de tratamentos nos diferentes mecanismos de valores ausentes, principalmente no mecanismo NMAR. Sendo assim, um dos principais diferenciais dessa dissertação quando comparada com os trabalhos supracitados, consiste na análise do comportamento de tratamentos de valores ausentes considerando os três mecanismos de ausência existentes na literatura da área.

4.3 MÉTODOS DE IMPUTAÇÃO BIOINSPIRADOS

Após a apresentação de trabalhos que retratam sobre tratamento de valores ausentes com enfoque no tipo de mecanismo utilizado em seus experimentos, esta seção trata especificamente

de trabalhos que utilizam tratamentos de valores ausentes, apresentando as principais diferenças entre eles e esta dissertação.

A utilização de modelos bioinspirados para tratamento de valores ausentes ainda é incipiente. Alguns trabalhos utilizam esses modelos como auxiliador de outras técnicas de forma a melhorar a convergência do método, os chamados tratamentos híbridos; poucos são os estudos que utilizam o modelo evolucionário para a imputação propriamente dita. Além disso, os trabalhos encontrados possuem domínios de aplicação diversos, desde desempenho de equipamentos ao desenvolvimento de padrões agrícolas, mostrando a capacidade de adaptação ao contexto do problema que os modelos bioinspirados possuem.

Sobre tratamentos híbridos, a utilização de algoritmos genéticos para aumentar a convergência das redes neurais na etapa de treino é antiga. Os trabalhos ABDELLA & MARWALA (2005.a, 2005.b) são considerados os precursores de tal abordagem no âmbito da imputação de dados. Outro trabalho que faz aplicação desse tipo de método é o de DHLAMINI, NELWAMONDO e MARWALA (2007), onde os autores agregam a uma rede neural do tipo *autoenconder* os algoritmos evolucionário AG e PSO. Ambos os métodos são avaliados e comparados para estimar os valores ausentes de buchas de alta tensão. Logo, os valores precisam ser de acordo com as medidas padrões do equipamento e são mensuradas baseada na acurácia dos modelos gerados por cada tratamento. Pela característica do problema, as bases podem ser consideradas de acordo com o mecanismo aleatório de ausência (MAR), porém nada é referenciado no trabalho. A conclusão dos autores sobre os métodos é que ambos geram resultados satisfatórios para o problema de domínio, considerando o algoritmo PSO melhor, pois executa o processo em um menor tempo e alcança o mesmo nível de acurácia do AG. Apenas esses algoritmos são testados e comparados, ou seja, não usam os métodos tradicionais de TVA para solucionar o problema e comparar com os métodos desenvolvidos, além de realizar os experimentos considerando um mecanismo. O principal diferencial recaí no fato de que este trabalho utilizou apenas um mecanismo de valor ausente para analisar os métodos.

Outro exemplo de uma proposta hibrida é o trabalho de ANDRADE, SILVA e HRUSCHKA (2009). Nele, foi implementado um algoritmo evolucionário baseado em agrupamento para imputação dos dados, o qual foi submetido a uma avaliação para problemas de classificação com cinco bases do repositório do UCI (ASUNCION e NEWMAN, 2007). No trabalho é mencionado que essas bases são completas e seus valores ausentes são inseridos de

acordo com uma distribuição completamente aleatória (MCAR); e além disso, considera apenas atributos numéricos pois, o algoritmo utiliza-se da distância Euclidiana para calcular a dissimilaridade entre agrupamentos. De modo geral, o método proposto é um combinado do agrupamento por KNN e algoritmo genético; a avaliação ocorreu em comparação com outros três métodos de imputação por agrupamento de dados, o KNN (BRÁS; MENEZES, 2007), SKNN (KIM; KIM; YI, 2004) e IKNN (TROYANSKAYA et al., 2001). Assim, percebe-se também que o trabalho utiliza apenas um dos mecanismos para avaliar seu tratamento bioinspirado.

Em (SILVA; HRUSCHKA 2013), confrontam seu tratamento bioinspirado híbrido com outros métodos de tratamento e considera dois mecanismos de valores ausentes, o MAR e MCAR. Nesse trabalho foram utilizados apenas métodos de imputação baseado em agrupamento de dados, e estes tiveram seus valores analisados segundo sua acurácia para seis classificadores de acordo com seis bases de dados diferentes. Logo, percebe-se que houve um extenso processo de experimentação, contudo analisando dois mecanismos de ausência.

Os algoritmos biosinpirados mais recorrentes em tratamentos híbridos são o AG e o PSO. Todavia, alguns exemplos vão além desses, como os de VERONEZE et al. (2011) e de FRANÇA, COELHO e VON ZUBEN (2013). Nesses dois trabalhos, uma técnica chamada de bio-clusterização é utilizada. Onde no primeiro trabalho foi aplicado o método em uma base de expressão genética completa, onde os mecanismos de ausência são simulados de acordo com os padrões MCAR, MAR e NMAR. O trabalho investiga a qualidade da imputação gerada pelo método em comparação com outros dois métodos de imputação baseados em agrupamento utilizando a métrica RMSE. Vale ressaltar que os testes são realizados em apenas uma base de dados, o que o diferencia do desta dissertação que aplica os mecanismo à sete bases diferentes. Já o segundo trabalho é considerado uma atualização do primeiro, havendo a melhoria de pontos fracos da observados na primeira abordagem, principalmente quanto a forma de como o modelo anterior estimava os valores ausentes. Logo, o trabalho também recai nas mesmas métricas utilizadas e também utiliza apenas uma base para comparação do comportamento do seu algoritmo.

Acerca dos modelos bioinspirados para imputação propriamente dita, estudos envolvendo AG destacam-se por sua contribuição à um determinado domínio de aplicação como em FIGUEROA GARCÍA, KALENATIC e LOPEZ BELLO (2008), e FIGUEROA GARCÍA, KALENATIC e LOPEZ BELLO (2010). Em tais pesquisas, os autores utilizaram medidas

estatísticas, como a matriz de covariância e a função de auto-correlação de instâncias sem valores ausentes, onde se realizava o processo de imputação e depois calculava as mesmas medidas, a fim de verificar a distância entre elas. Tal abordagem também pode ser observada em FIGUEROA GARCÍA, KALENATIC e LOPEZ BELLO (2011). Nesses trabalhos, apesar de não estar explicitamente descrito, infere-se que as variáveis obedecem o mecanismo de ausência MCAR. Por tanto, os trabalhos utilizam apenas um mecanismo de ausência.

Outra pesquisa relevante a qual utiliza-se de um algoritmo genético é o de (PATIL, 2010), implementado para tratar da problemática sob o ponto de vista da imputação múltipla. Tal trabalho possui pontos falhos em sua essência, como uma não explicitação acerca do mecanismos de ausência das bases utilizadas; as métricas também possuem pouca explicações de como foram mensuradas e tendem a confundir o leitor em relação as conclusões mencionadas pelo autor. Contudo, considera-se que o experimento foi gerado de acordo com o mecanismo MAR, pois há uma dependência entre os atributos removidos, sendo este o diferencial entre esta dissertação e a referida pesquisa.

Ainda sob o ponto de vista de trabalhos que utilizam AG, a aplicação de (AZADEH et al., 2013) também merece ser mencionada. O domínio em questão é a projeção de experimentos agrícolas, onde são comparados três método de imputação evolucionários (AG, PSO e RNA) e um método de regressão. Os experimentos foram realizados em cinco tratamentos e cinco blocos do domínio. Todavia, as tabelas, como são chamadas pelos autores, não possuem valores ausentes. A fim de simular tal problema, foi realizada uma abordagem diferente das dos mecanismos usualmente conhecidos: é sempre dito que dois valores ausentes são encontrados em uma instância da tabela, onde o primeiro valor é arbitrariamente escolhido e o segundo é calculado de acordo com o primeiro valor simulado, tal procedimento se repete até todas as instâncias da tabela conterem VA. Por meio da modelagem, pode-se atribuir que as bases geradas fazem alusão ao método aleatório (MAR), uma vez que o segundo é mensurado a partir do primeiro valor. Portanto, o trabalho efetuou sua análise em apenas um mecanismo.

Acerca das abordagens acima citadas, observa-se a capacidade de aplicar métodos bioinspirados tanto para imputação de dados propriamente dita, quanto para auxiliar outros métodos, conseguindo ainda adaptar o modelo de acordo com o domínio de aplicação. Todos os trabalhos mostram que a utilização de métodos evolucionários geram resultados promissores, havendo ampla possibilidade para pesquisas futuras, como em comparações com propostas

tradicionais de TVA e o comportamento em diferentes mecanismos. Tal fato corrobora a utilização de algoritmos bioinspirados na presente dissertação, que tem como principal diferencial a aplicação um método bioinspirado para imputação dados, comparando-o com outros métodos tradicionais da área e analisando seu comportamento para os diferentes mecanismos de ausência de dados existentes.

4.4 SÍNTESE DO CAPÍTULO

Neste capítulo foram mostrados trabalhos na literatura de tratamento de valores ausentes, os quais possuem a utilização de um ou mais mecanismo durante seus experimentos; utilizam algoritmos bioinspirados para estimar os valores removidos das bases de acordo com algum mecanismo; e trabalhos que geram bases ausentes baseadas nesses mecanismos.

Os principais diferenciais do presente trabalho quando comparado com esses, está no fato de realizar uma análise de um algoritmo bioinspirado pouco explorado na literatura sob o ponto de vista dos três mecanismos de ausência dos dados e, ainda o compara diferentes configurações dele próprio e com outros métodos de tratamento conhecidos na área.

Maiores informações sobre como cada um desses sistemas, sua codificação e diferenciais estão descritos nos próximos capítulos.

5 TRATAMENTO DE VALORES AUSENTES UTILIZANDO ENXAME DE PARTÍCULAS

5.1 CONSIDERAÇÕES INICIAIS

Após apontar as diferenças entre este trabalho e outros encontrados na literatura de tratamento de valores ausentes, passa-se a explicação do funcionamento e desenvolvimento do método utilizando enxame de partículas para estimar os valores ausentes em uma determinada base de dados. Neste capítulo são elucidados os seguintes pontos pertinentes ao método: modelagem da partícula, função de aptidão, cálculos de velocidade e inércia.

5.2 TRATAMENTO DE VALORES AUSENTES UTILIZANDO ENXAME DE PARTÍCULAS

Inspirado no algoritmo de otimização por enxame de partículas apresentado no Capítulo 3 desta dissertação, o método para tratamento de valores ausentes apresentado por este trabalho foi modelado. Porém, algumas modificações foram feitas para assim adaptar o algoritmo ao problema.

O PSO foi selecionado entre outros métodos bioinspirados por ainda não ter sido detalhadamente analisado na área de tratamento de valores ausentes e por ter sua implementação facilitada pela sua estrutura, a qual conta com uma menor quantidade de parâmetros e funções a serem codificadas para seu pleno funcionamento.

Além dessa escolha, também foi necessário verificar a melhor estratégia para o desenvolvimento do mesmo. Para isso pensou-se que desde a manipulação das bases até a implementação do método de imputação seria utilizada a ferramenta WEKA e a linguagem JAVA. Uma vez que a primeira é utilizada no meio acadêmico de Aprendizado de Máquina e possui implementações prontas de tarefas necessárias para esta dissertação, e.g. manipulação de

bases; e a segunda possibilita a utilização do paradigma orientado à objetos, o que facilitou a codificação do método.

Durante o desenvolvimento do algoritmo para a estimação dos valores ausentes, verificou-se que com o tratamento desse problema há uma melhora na qualidade dos dados de uma base, já que com os valores estimados é possível melhorar o reconhecimento de padrões se comparado quando não havia valores para os exemplos. Com base nisso, buscou-se na literatura exemplos de tarefas de Mineração de Dados que tinham dificuldade de serem realizadas caso a base contivesse dados faltosos. Como resultado, observou-se que a tarefa de classificação é a mais analisada nos trabalhos presentes na literatura da área. Isso ocorre pois para se classificar uma determinada classe é preciso a maior quantidade de padrões possíveis para que novas entradas sejam reconhecidas pelo classificador, com isso, caso haja valores ausentes na base alguns padrões podem ter seu reconhecimento comprometido. Com base em tais observações, o algoritmo de tratamento por otimização por enxame de partícula foi modelado para tarefas de classificação.

5.2.1 MODELAGEM DA PARTÍCULA

O primeiro passo para a utilização do método de otimização por enxame de partículas como tratamento de valores ausentes consiste em adaptar as suas partículas para o problema. A forma pelas quais as partículas foram modeladas para esta dissertação seguem o padrão representado pela Figura 2, onde podemos observar a ocorrência de valores ausentes em uma base de dados didática, a formação da chamada lista com os valores de domínio e da partícula propriamente dita. Todos esses passos são explicados a seguir.

Base com valores ausentes

Sexo	Idade	Gravidez	Classe
Feminino	?	Positivo	Preferencial
Masculino	75	Negativo	Não-preferencial
?	30	Negativo	Não-preferencial
?	27	Positivo	Preferencial
Masculino	19	?	Não-preferencial
Feminino	20	Positivo	Preferencial

Lista de valores de domínio

Classe	Atributo	Domínio
Preferencial	Idade	{75,27,20}
	Sexo	{Feminino}
Não-preferencial	Gravidez	{Negativo}
	Sexo	{Masculino}

Partícula - 0

1-2	4-0	3-0	5-0

Partícula - 1

1-0	4-0	3-0	5-0

Figura 2. Procedimento de criação da lista de valores de domínio e da particular do algoritmo.

Ao iniciar o algoritmo de tratamento, ocorre a leitura e reconhecimento de cada um dos exemplos contidos na base, incluindo os faltosos. Logo após é realizado o mapeamento das classes e atributos que possuem valores ausentes. Além disso, também é feito o armazenamento do índice do atributo ausente pertencente a uma determinada classe pois, dessa forma é possível evitar que um determinado padrão de um atributo de uma classe seja copiado para outra a qual não pertença, como por exemplo, se uma pessoa está grávida, ela necessariamente pertence ao sexo feminino, o que exclui a possibilidade de estimar o valor de sexo masculino caso este atributo esteja ausente e o atributo gravidez esteja com o valor positivo.

Em posse do mapeamento de todos os valores ausentes, o algoritmo passa a montar a lista com os valores de domínio para cada um dos atributos das classes ausentes. Tais valores são usados pelo enxame de partículas como possíveis valores estimados para o exemplo ausente. Tal procedimento ocorre da seguinte maneira: busca-se por todos os valores presentes na base (“não-

ausentes") os quais pertençam ao atributo e a classe ausente. Com isso, o algoritmo cria uma lista para cada atributo ausente a qual é composta por todos os valores observáveis na base para aquele determinado atributo. Tal distribuição vale para valores numéricos e nominais, os quais são tratados da mesma forma, ou seja, mesmo se o atributo seja contínuo apenas os valores que aparecerem na base são usados para imputar os ausentes.

Depois de criadas as listas com os valores de domínio para cada atributo ausente o método passa para a modelagem da partícula, a qual é formada por outra lista, dessa vez contendo um par "índice-valor" de todos os valores ausentes da base, onde o índice corresponde a linha do exemplo ausente e o valor é o índice correspondente na lista de valores de domínio previamente encontrados para o atributo em questão, ou seja, caso esteja aparecendo a dupla "22-4" sabe-se que esse valor ausente encontra-se na linha 22 e o valor de índice 4 na lista de domínio foi o valor estimado. Essa lista da partícula está dividida pelas classes e atributos ausentes já mapeados, sendo assim, o primeiro item dessa lista pertence ao primeiro atributo ausente contido em uma primeira classe também ausente, e assim é formada uma partícula do enxame.

Esses procedimentos correspondem aos passos realizados pelo método para modelar a partícula. Dessa forma, cada uma contém todos os valores ausentes na base e o método sabe exatamente a localização de cada um desses valores. Vale ressaltar que o primeiro enxame, ou população inicial do método, é criado com valores de índices aleatórios para o par "valor" da lista da partícula, onde o número de partículas é determinado pelo usuário – a quantidade utilizada nessa dissertação encontra-se descrita no Capítulo 6.

5.2.2 FUNÇÃO DE APITIDÃO

Uma vez que o algoritmo de tratamento foi modelado para melhorar a qualidade dos resultados da tarefa de classificação quando houvesse a ocorrência de valores ausentes em uma determinada base, e a partícula do PSO modelada para que contivesse todos os valores estimados para os exemplos ausentes. Escolheu-se que a aptidão do algoritmo seria a própria classificação da base, onde cada partícula gera uma base completa, a qual é formada pelos valores observáveis na base e os estimados pela partícula, e é submetida à um classificador em cada iteração do algoritmo. Dessa forma, a base é avaliada de acordo com a finalidade do algoritmo de melhorar a

qualidade dos dados, onde quanto maior o valor da acurácia encontrada pela combinação dos valores estimados e dos valores presentes na base, melhor a partícula.

Tendo definido a tarefa de Mineração de Dados para análise dos dados com valores ausentes, a modelagem da partícula e a função de aptidão dela, decidiu-se por observar o grau de adaptabilidade do algoritmo bioinspirado para estimação dos valores faltosos na classificação. Sendo assim, foram implementados três classificadores utilizados como função de aptidão do PSO: Naïve Bayes, C4.5 e n-vizinho mais próximo (do inglês *K Nearest Neighbor – KNN*) com 3 vizinhos - todos implementados com o suporte da ferramenta WEKA (HALL et al., 2009). Tal escolha ocorreu de forma a se obter uma amostra de cada tipo de classificador baseado na tipologia apresentada por (LUENGO, 2011). Tais classificadores são exemplos de modelos aproximados, aprendizado por indução de regras e *lazy learning*, respectivamente, onde o primeiro aplica probabilidade com base na aplicação da teoria de Bayes com suposições de independência fortes; o segundo gera regras em forma de árvores; e o terceiro classificador aproxima uma função de um ou mais pontos (vizinhos) a fim de agrupar os mais similares a este primeiro (FACELI et al., 2011).

Em suma, cada partícula origina uma base completa a qual é submetida aos classificadores supracitados, por meio do 10-*fold-crossvalidation*, gerando um valor de acurácia, sendo este armazenado no algoritmo como função de *fitness* de cada partícula. Em posse desse valor, o algoritmo pode realizar seus cálculos de atualização da velocidade, posição e inércia.

5.2.3 CÁLCULO DA VELOCIDADE E INÉRCIA

Afim de balancear a busca pelo espaço local e global, optou-se por acrescentar mais uma variável na Eq. (5) chamada de peso de inércia (*w)*, dando origem a nova equação de velocidade, adaptada de Shi & Eberhart (1998) e representada na Eq. (7).

$$V(t+1) = w * V(t) + b1 * Rnd * \left(\mathrm{mLocal} - \mathrm{X(t)}\right) + b2 * Rnd * \left(mGlobal - X(t)\right) \quad (7)$$

Onde *w* é chamado de peso de inércia aleatório, apresentado pela Eq. (8), a qual segundo (BANSAL et al., 2011), promove uma melhor eficiência para o enxame.

$$w = 0.5 + \frac{Rand()}{2} \tag{8}$$

onde *Rand()* é uma função aleatória a qual retorna um valor entre 0.0 e 1.0.

Com base nessas duas equações, o algoritmo de imputação por enxame de partículas calcula a sua velocidade para modificar a posição, e o valor estimado de cada dimensão da partícula. A velocidade está relacionada com a distância atual da partícula para a melhor posição já encontrada. Sendo assim, no algoritmo apresentado nesta dissertação, a velocidade influencia na estimação do próximo valor imputado, o qual é um número aleatório da distância atualizada baseado na lista de valores de domínio para cada valor ausente, fazendo com que o algoritmo possa encontrar uma combinação ainda melhor do que a anterior. Por exemplo, se a velocidade atualizada retorna um valor igual à 4, o próximo valor estimado será um valor aleatório no intervalo entre 0 e 3 da lista de domínio do valor ausente de determinada dimensão da partícula. Dessa forma, cada valor ausente é caracterizado como sendo uma dimensão da partícula e possui sua própria velocidade.

5.3 SÍNTESE DO CAPÍTULO

Esse capítulo apresentou como o algoritmo de imputação de dados utilizando a otimização por enxame de partícula foi implementado. Foram descritos a estratégia de modelagem da partícula, a qual utiliza apenas os valores ausentes da base; a escolha da função de aptidão, tendo sido selecionada a acurácia obtida na tarefa de classificação; suas diferentes configurações utilizadas nos experimentos, os classificadores Naïve Bayes, C4.5 e 3-NN; e por ultimo, como o cálculo da velocidade foi modificado e adaptado para o problema de valores ausentes a fim de melhorar a eficiência do algoritmo.

6 CONFIGURAÇÕES E REALIZAÇÃO DOS EXPERIMENTOS

6.1 CONSIDERAÇÕES INICIAIS

Após apresentar o funcionamento do algoritmo de imputação por PSO, o mesmo precisa ser validado. Para tal, é necessária a realização de experimentos, tendo as origens das bases usadas e os procedimentos de sua execução explicados durante este Capítulo. Além dessas informações, também são mostradas as configurações do método apresentado nesta dissertação, dos métodos com os quais é comparado e a métrica de avaliação utilizada entre eles.

6.2 CONFIGURAÇÕES DOS EXPERIMENTOS

A fim de observar o comportamento do método de imputação por PSO, foram geradas 63 bases de dados sintéticas com seus valores ausentes padronizados de acordo com os mecanismos de ausência de dados MAR, MCAR e NMAR. Para alcançar tal quantidade, foram selecionadas 7 bases de dados do repositório de aprendizado de máquina do UCI (ASUNCION e NEWMAN, 2007), escolhidas com o intuito de representar as diferentes configurações das bases existentes e para observar a capacidade de adaptação do método de imputação, são elas: contraceptive, glass, íris, lymphography, tic-tac-toe, vertebral-column e yeast. As suas propriedades como quantidade de instâncias, atributos e classes são mostrados na Tabela 6.1.

Tabela 6.1. Propriedades das bases de dados selecionadas para a realização dos experimentos.

Nome	**Instâncias**	**Atributos**	**Classe**
Contraceptive	1473	9	3
Glass	214	10	6
Iris	150	4	3

Lymphographic	148	18	4
Tic-tac-toe	958	9	2
Vertebral Column	1484	8	10
Yeast	310	6	2

A partir das bases supracitadas foram geradas regras de acordo com os mecanismos de ausência de dados para cada uma delas utilizando a teoria apresentada no Capítulo 2.

Não bastante a análise baseada nos mecanismos de ausência, também decidiu-se analisar o comportamento do método para diferentes quantidades de valores ausentes em um ou mais de seus atributos. Assim, para cada uma das bases de um determinado mecanismo, foram removidas as quantidades de 15%, 30% e 45%.

A escolha dos atributos deu-se pela seleção do qual possuía o maior valor de ganho para a base e o mais correlacionado com este. Com isso os valores removidos afetam a qualidade dos dados de uma base, fazendo com que os valores imputados também apresentem um efeito sobre tal observação.

Depois de aplicar tais regras nas bases originais, 63 versões sintéticas foram geradas. Com isso, outras propriedades foram adicionadas à elas, como porcentagem de valores ausentes e porcentagens de instâncias com valores ausentes. Em suma, essas características adicionais são representadas na Tabela 6.2, onde o nome das bases sintéticas seguem o seguinte padrão: nomeDaBase_mecanismo_quantidadeAusente para facilitar a diferenciação entre elas. As demais bases estão listadas no APÊNDICE B desta dissertação.

Tabela 6.2. Características das bases sintéticas geradas.

Nome	**%V.A.**	**% Inst. V.A.**
Glass_MCAR_15	2,99	27,10
Glass_MAR_30	0,56	5,60
Glass_NMAR_45	2,10	21,02
Lymphographic_MCAR_15	0,74	7,47
Lymphographic_MAR_30	3,12	53,37

Lymphographic_NMAR_45	1,92	36,48
Contraceptive_MCAR_15	2,98	27,76
Contraceptive_MAR_30	3,57	35,77
Contraceptive_NMAR_45	5,09	50,91
Iris_MCAR_15	0,93	4,66
Iris_MAR_30	1,73	8,66
Iris_NMAR_45	7,33	36,66
Tic-tac-toe_MCAR_15	1,38	12,42
Tic-tac-toe_MAR_30	5,87	49,26
Tic-tac-toe_NMAR_45	8,31	65,13
Yeast_MCAR_15	0,17	1,75
Yeast_MAR_30	0,42	4,24
Yeast_NMAR_45	1,39	13,94
Vertebral Column_MCAR_15	4,23	27,74
Vertebral Column_MAR_30	2,99	20,96
Vertebral Column_NMAR_45	5,76	40,32

Após a apresentação das bases utilizadas para os experimentos, sua geração e configurações, foca-se em como o método de imputação por PSO procedeu para executa-las.

Primeiramente, buscou-se uma validação do PSO da forma mais ampla possível. Pensando nisso, o método foi desenvolvido com diferentes funções de aptidão pois assim, o comportamento sobre essa escolha também seria analisada. Como a aptidão do algoritmo está relacionada à tarefa de classificação, os classificadores selecionados foram: Naïve Bayes, C4.5 e 3-NN, implementados de acordo com o explicado na sessão 5.2.2 desta dissertação.

Em seguida, já que o método possui uma inicialização estocástica, decidiu-se executar cada uma das configurações da função de aptidão do PSO 10 vezes para cada base sintética, totalizando 630 experimentos para cada uma.

6.3 CONFIGURAÇÕES DOS MÉTODOS DE IMPUTAÇÃO UTILIZADOS

Em posse das informações sobre as bases de dados utilizadas nos experimentos e de como procedeu a execução delas pelo método de imputação por PSO, apresenta-se como tal método foi configurado para ser avaliado. Como mencionado, o tratamento estudado por esta dissertação foi desenvolvido com três funções de aptidão diferentes. Logo, além das configurações do enxame de partículas, como quantidade de partículas, iterações e função de inércia, também são necessários mostrar as dos classificadores usados no método. A Tabela 6.3, Tabela 6.4 e Tabela 6.5 mostram, respectivamente o resumo das configurações do PSO e dos classificadores 3-NN e C4.5. Para o classificador Naïve Bayes não houve escolhas de parâmetros diferentes do padrão oferecido pela ferramenta WEKA.

Tabela 6.3. Configuração do tratamento de valores ausentes por enxame de partículas.

Quantidade de partículas	50	
Quantidade de iterações	100	
Inércia	$w = 0.5 + \frac{Rand()}{2}$	(8)

Tabela 6.4. Configuração do classificador 3-NN.

Número de vizinhos	3
Algoritmo de busca de vizinhos	Distância euclidiana

Tabela 6.5. Configuração do classificador C4.5.

Fator de confidencia	0,25
Número mínimo de instâncias por folha	2
Número de *folds*	3

Além das configurações do tratamento de valores ausentes por PSO e de seus classificadores, durante os experimentos foram executados outros dois métodos de tratamento conhecidos da área, o baseado em máquina de vetores de suporte, chamado SVMImpute e outro baseado no vizinho mais próximo, chamado de KNNImpute. As 63 bases sintéticas foram importadas para a ferramenta KEEL, onde lá foram selecionados os método acima citados com as seguintes configurações.

Tabela 6.6. Configurações dos métodos de imputação KNNImpute e SVMImpute.

KNNImpute	
Número de vizinhos	10
SVMImpute	
Tipo de regressão	Epsilon
Tipo de kernel	RBF

6.4 FUNÇÃO DE AVALIAÇÃO

Depois de apresentar as origens das bases usadas, como foram configuradas para serem usadas nos experimentos, mostrar as configurações dos métodos de imputação por PSO, SVMImpute e KNNImpute, tais testes precisam passar por uma função para serem avaliados. Por se tratar de experimentos utilizando bases com valores ausentes controladamente gerados, buscou-se na literatura qual a função utilizada para esse tipo de base. Assim, selecionou-se o cálculo da raiz quadrada do erro quadrático médio (RMSE – do inglês, *Root mean square error*) entre o valor imputado por cada um dos tratamentos e o valor original para cada base, descrita pela Eq.(9).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \hat{e}_i)^2} \quad (9)$$

6.5 SÍNTESE DO CAPÍTULO

Durante este capítulo informações como quais bases foram utilizadas nos experimentos, como tiveram seus valores removidos e como o tratamento por PSO as processou foram apresentados. Além dessas, as configurações dos métodos de tratamentos usadas e dos classificadores implementados pelo enxame de partículas, também foram mostradas. Por fim, a função de avaliação para os experimentos realizados nesta dissertação foi explicada.

Com base nas informações descritas por este capítulo, os experimentos foram realizados gerando os resultados mostrados e analisados conforme o próximo capítulo descreverá.

7 RESULTADOS DOS EXPERIMENTOS

7.1 CONSIDERAÇÕES INICIAIS

Neste capítulo, os resultados dos experimentos realizados com os métodos de imputação são mostrados em forma de gráficos gerados a partir de cada uma das configurações mostradas no capítulo anterior. Além disso, são analisados aspectos similares de cada experimento, gerando uma forma mais geral de visualizar o comportamento do tratamento de VA por PSO e de compará-lo com métodos já usados na área e com diferentes configurações dele mesmo.

7.2 RESULTADOS E ANÁLISE DOS EXPERIMENTOS

A partir de 63 experimentos realizados, foi possível observar o comportamento de cada um dos métodos de imputação utilizados. Mais especificamente, o desempenho de três diferentes configurações do algoritmo de imputação com PSO puderam ser analisadas e compradas entre métodos conhecidos da literatura da área e entre si. Os resultados apresentados nesta seção para o método de imputação por PSO, são formados pelo menor valor de RMSE encontrado por cada configuração do método em questão entre todas as 10 execuções de cada experimento.

Dentre os resultados gerados, os valores obtidos para a base vertebral-column foram os que mais divergiram dos demais. Para essa base, cada método obteve o mesmo valor de RMSE para todos os mecanismos e porcentagens de valores ausentes, como mostra a Figura 3.

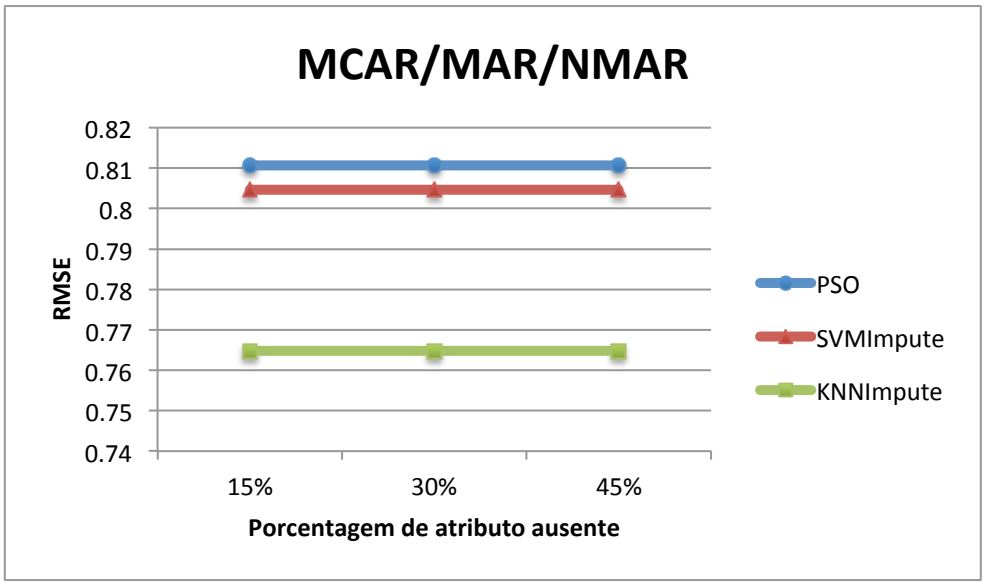


Figura 3. Gráfico da base vertebral-column.

No gráfico acima, as três configurações para o algoritmo de imputação por PSO foram agregadas à uma única legenda pois obtiveram o mesmo valor de RMSE. Ademais, como todos os métodos alcançaram os mesmos valores, não houve alternância entre os que obtiveram os melhores resultados. Assim, o KNNImpute pôde ser considerado o melhor para essa base quando comparado com os outros.

Ao contrário da base vertebral-column, nas demais observou-se que havia uma variação entre os métodos que alcançaram os valores mais baixos de RMSE quando analisado os mecanismos de ausência de dados e a porcentagem de ausência nos atributos. Decerto, essa característica corrobora o fato de não haver um método de imputação geral, que seja ótimo para qualquer base, mecanismo ou quantidade de VA. A Figura 4, Figura 5 e Figura 6 são exemplos desta observação, tais resultados foram gerados a partir das bases tic-tac-toe, contraceptive e íris, respectivamente.

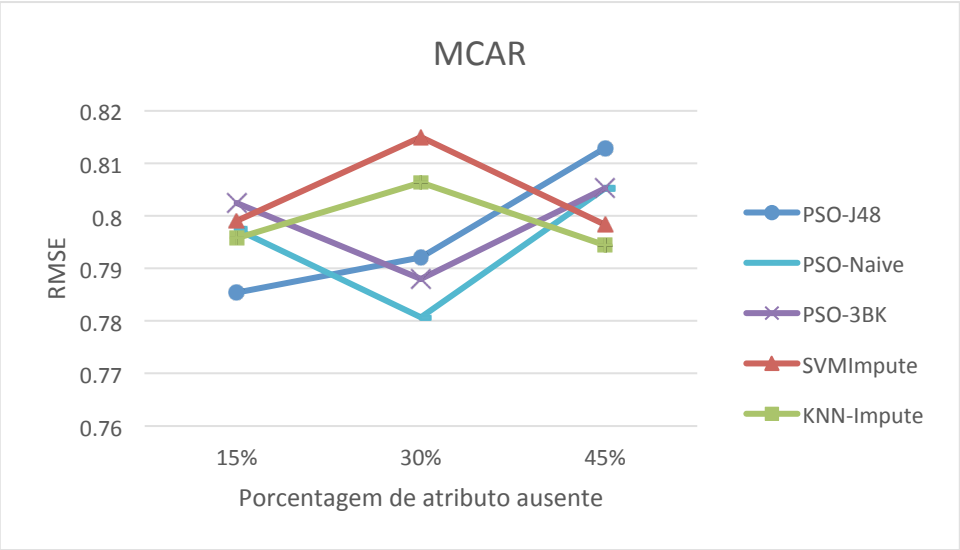


Figura 4. Gráfico dos resultados do mecanismo MCAR para a base tic-tac-toe.

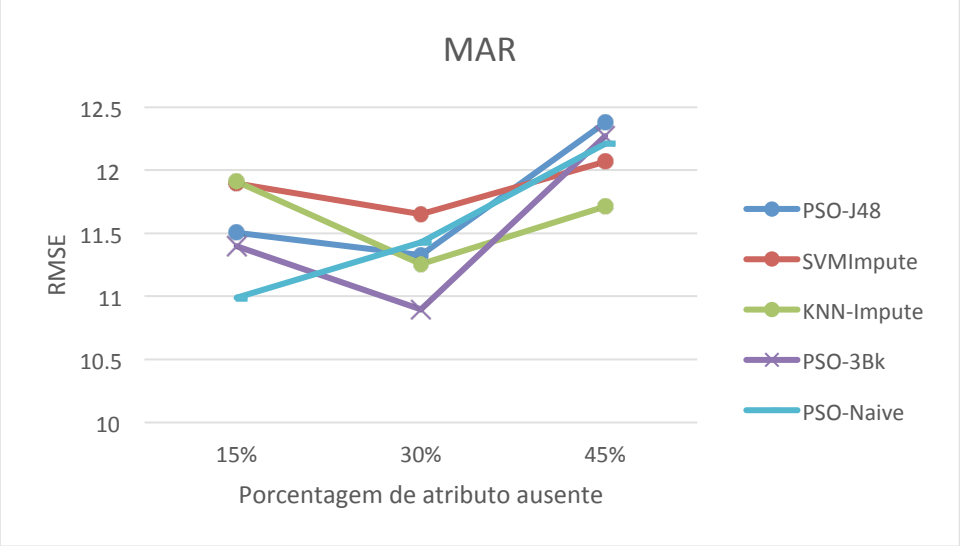


Figura 5. Gráfico dos resultados do mecanismo MAR para a base contraceptive.

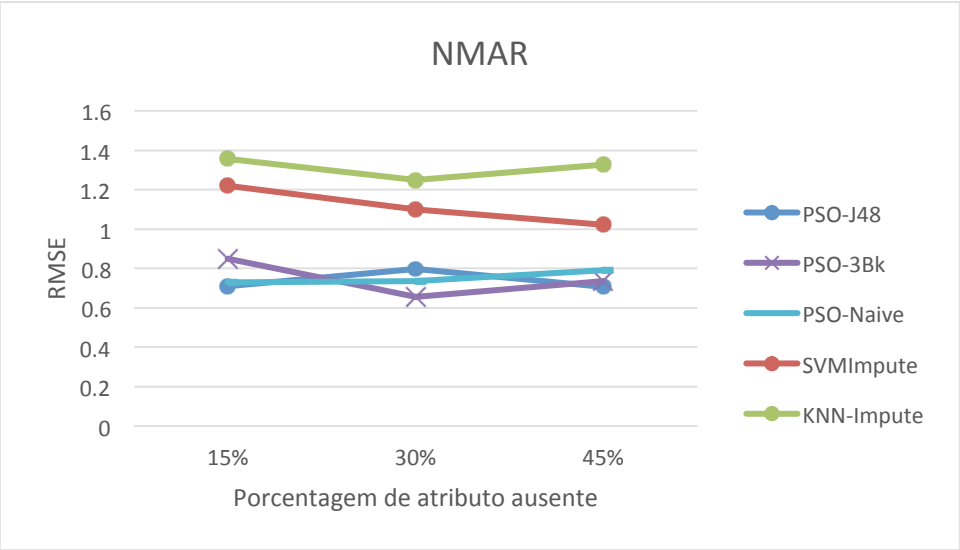


Figura 6. Gráfico dos resultados do mecanismo NMAR para a base íris.

Os resultados dos demais experimentos para cada base de dados selecionadas são listados no APÊNDICE A deste documento.

Além das comparações diretas entre os métodos, também foram gerados gráficos mais gerais dos resultados, com o intuito de analisar o comportamento dos tratamentos sobre diferentes aspectos, como por exemplo, o método que obteve a maior quantidade de melhores resultados em geral, para cada porcentagem de atributo ausente e mecanismo de ausência de dados. Para essas análises, o método de imputação por PSO teve os valores de suas configurações agregados por meio da média entre eles, sendo assim possível realizar inferências sobre o método como um todo.

Considerando tais aspectos, a Figura 7 ilustra o resultado da análise do comportamento dos métodos de imputação tendo em vista a porcentagem de atributo ausente.

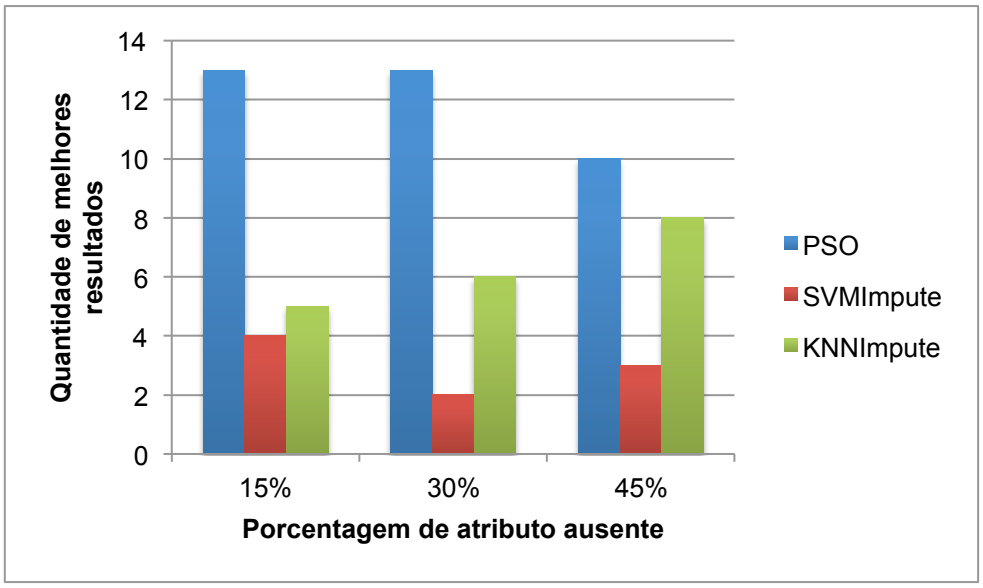


Figura 7. Gráfico dos melhores resultados obtidos pelos métodos de tratamento de VA considerando diferentes porcentagens de ausência no atributo.

Ao analisar a figura acima, percebe-se que o algoritmo utilizando PSO obteve a maior quantidade de melhores resultados para as diferentes porcentagens de atributo ausente, como um todo, considerando todas as bases dos experimentos. Além disso, pode-se observar que quanto maior a quantidade de VA na base, menor a performance do tratamento por PSO, fato o qual ocorre de forma inversa se analisado o tratamento KNNImpute. Porém, mesmo melhorando sua performance, não supera o por enxame de partículas.

Além deste primeiro aspecto, também foram observados os comportamentos dos métodos para cada um dos mecanismos de ausência de dados, tendo seus resultados ilustrados pela Figura 8.

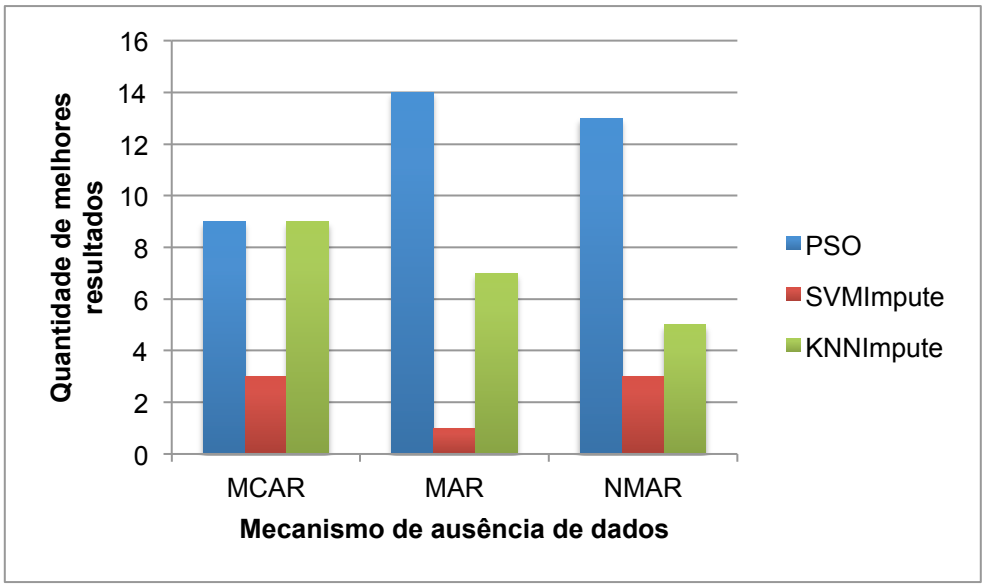


Figura 8. Gráfico dos melhores resultados obtidos pelos métodos de tratamento de VA considerando os mecanismos de ausência de dados.

Por meio do gráfico, observa-se que o tratamento por PSO ainda mantém uma certa superioridade em relação a quantidade de melhores resultados encontrados em relação aos demais, com exceção para o mecanismo MCAR onde foram obtido a mesma quantidade que o KNNImpute. Vale ressaltar que esse tipo de análise, para uma quantidade considerável de bases com VA, dificilmente é retratado na literatura da área por conta de que o mecanismo NMAR ser considerado complexo e pouco utilizado para geração de bases sintéticas.

Além dos gráficos anteriormente gerados, decidiu-se analisar de forma mais ampla os métodos de tratamento de VA, dando origem a Figura 9. Onde nela nota-se que o tratamento por PSO permanece obtendo os melhores resultados, refletindo de forma mais evidente o que as análises anteriores também demostraram.

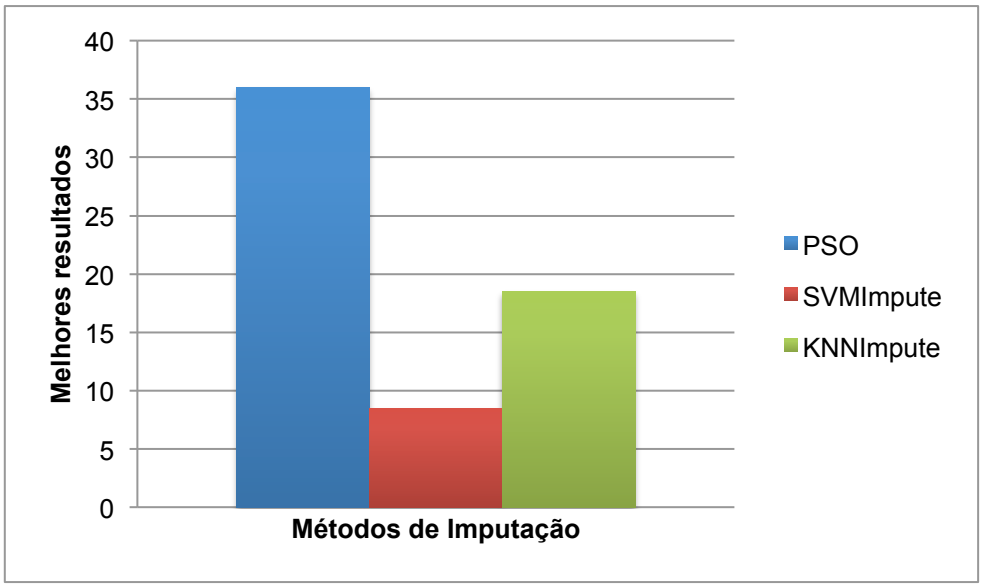


Figura 9. Gráfico dos melhores resultados obtidos pelos métodos de tratamento de VA em geral.

Vale ressaltar que para o cálculo dos melhores resultados foi considerado qual método obteve o menor valor de RMSE para determinada base de um determinado mecanismo e percentual de atributo ausente. Dos 63 experimentos realizados, foi encontrado um valor igual para os tratamentos SVMImpute e KNNImpute na base yeast, mecanismo MAR para 15% de atributo ausente. Com isso, ambos os métodos obtiveram o melhor resultado para esse experimento, o que acaba por refletir nos gráficos da Figura 7 e da Figura 8, as quais totalizaram 64 melhores resultados. O cálculo realizado para o gráfico da Figura 9 considera a quantidade de 0,5 para cada um dos métodos, o que não afeta a quantidade total de resultados mas sim na individual deles, contabilizando 8,5 e 18,5 para o SVMImpute e KNNImpute, respectivamente.

Após a apresentação dos resultados obtidos nos experimentos realizados, consegue-se afirmar que o tratamento de VA por PSO alcança resultados que favorecem sua utilização na área. Além disso, também foi possível ressaltar que mesmo com resultados superiores aos métodos KNNImpute e SVMImpute de modo geral, quando analisado cada um dos experimentos, verifica-se que não há um tratamento perfeito, pois na maioria dos casos, há a variação de qual método alcançou o melhor resultado.

7.3 SÍNTESE DO CAPÍTULO

Em suma, este capítulo apresentou os resultados dos experimentos realizados assim como as análises individuais e gerais. Utilizando os insumos gerados pelos resultados foi possível verificar o comportamento do método de tratamento de VA por PSO e compará-lo com métodos conhecidos da área e com diferentes configurações dele mesmo. Ademais, no próximo capítulo algumas conclusões sobre esta dissertação são apontados, assim como as dificuldades encontradas e trabalhos futuros.

8 CONCLUSÃO

O problema de valores ausentes é abordado em diversos trabalhos publicados nos últimos anos, os quais descrevem sua ocorrência danosa para análise dos dados, principalmente durante o processo de extração do conhecimento de base de dados.

Após uma análise de trabalhos os quais dissertassem sobre o tema publicados em periódicos e congressos e com diversas áreas de aplicação, verificou-se uma falta de padrão quanto a origem da base utilizada nos experimentos desses trabalhos.

Com base nessa lacuna a seguinte observação foi feita, as bases comumente utilizadas são oriundas de repositórios públicos como o UCI, porém essas podem ou não possuir valores ausentes em seus dados. Usualmente escolhem-se bases completas para depois remover, controladamente, valores dessa base para analisar o comportamento de um determinado método, originando as chamadas bases sintéticas. Para isso ocorrer, pesquisadores simulam alguns dos mecanismos de ausência existentes na literatura da área, dos quais os mais replicados são os MAR e MCAR, por serem os mais fáceis uma vez que a ausência dos dados está baseada nos valores presentes na base ou completamente aleatório. Porém, para uma análise mais extensa de como os tratamentos se comportam para esses mecanismos, torna-se mais interessante analisa-los para os três existentes: MAR, MCAR e NMAR, sendo raros os trabalhos que conseguem, efetivamente, realizar esse tipo de estudo, principalmente quando deseja-se utilizar uma quantidade maior de bases de dados.

Além da tendência de analisar o tratamento de valores ausentes sobre os aspectos dos mecanismos de ausência de dados, outro fator o qual vem se destacando nessa área é a utilização de métodos bioinspirados, seja auxiliando outros métodos, ou como método propriamente dito. Os resultados apresentados por recentes publicações mostram-se promissores para esses tipos de tratamentos porém, ainda muito incipiente.

Com base na observação realizada e na tendências descrita acima, a presente dissertação apresenta uma abordagem para auxiliar o desenvolvimento da área, a qual consiste na exploração de um algoritmo bioinspirado ainda pouco utilizado, o de otimização por enxame de partículas, ou PSO. A escolha desse método ocorreu pelo fato de este ter sua estrutura de partículas

facilmente remodeladas para o problema de tratamento de valores ausentes e possuir poucos parâmetros tanto de implementação quanto de configuração, o que facilita e instiga sua utilização.

Sendo assim, para promover um estudo relevante sobre o algoritmo, esta dissertação removeu valores de sete bases de dados de acordo com os mecanismos de ausência de dados MCAR, MAR e NMAR, e variando a quantidade de valores ausentes em três diferentes porcentagens, 15, 30 e 45% para um determinado atributo. Além disso, o PSO teve seus melhores resultados analisados segundo a métrica de RMSE e comparado com outros dois tratamentos, o KNNImpute e o SVMImpute. Um fato que deve ser ressaltado é que o PSO teve sua função de aptidão implementada para três classificadores diferentes, o C4.5, Naïve Bayes e 3-NN, gerando assim três diferentes tipos de PSO – tais configurações foram feitas com o intuito de analisar a adaptabilidade do algoritmo bioinspirado dada uma tarefa de mineração de dados, neste caso, a classificação.

Os resultados obtidos mostraram que, no geral, o método utilizando o PSO alcançou as menores diferenças entre os valores estimados e os valores originais na maioria dos experimentos. Entretanto, ocorreram algumas peculiaridades, como na base vertebral-column, onde apenas um dos três métodos obteve os melhores resultados para todos os mecanismos em todas as porcentagens, o KNNImpute. Além de tudo, esta dissertação também conseguiu corroborar o fato de não haver um tratamento ótimo para qualquer configuração base, uma vez que quando analisado cada uma individualmente percebe-se claramente a variação entre o tratamento que obteve o melhor resultado, fato este também mencionado em outros trabalhos da área.

Em posse dos resultados, essa dissertação alcançou seu principal objetivo, o de analisar o comportamento de um algoritmo de otimização por enxame de partículas adaptado para o problema de valores ausentes, utilizando diversas bases de dados e os diferentes mecanismos de ausência de dados. Além desses, também agrega as seguintes contribuições para área de tratamento de valores ausente:

- Simulou, efetivamente, os três mecanismos padrões de ausência: MCAR, MAR e NMAR; e
- Analisou e comparou o comportamento de diferentes configurações do PSO com outros dois algoritmos conhecidos da área.

Ademais, também foram encontrados algumas adversidades durante a realização desta dissertação, principalmente durante a etapa de testes, enumeradas a seguir:

- Falta de uma ferramenta que possua tratamentos de valores ausentes recentes já implementada.
- Ineficiência na exportação das bases geradas por métodos já implementados para um formato de fácil manipulação;
- Encontrar métricas diferentes do RMSE utilizado para avaliação dos métodos em bases sintéticas;
- Encontrar trabalhos na área para comparar resultados diretamente; e
- Utilizar bases que já venham de acordo com todos os mecanismos de ausência de dados.

Por fim, alguns trabalhos futuros são almejados para esta linha de pesquisa, são eles:

- Comparação com mais métodos de tratamento de valores ausentes;
- Aplicação de mais bases de dados;
- Aplicação das bases imputadas pelos outros métodos à tarefa de classificação para avaliar sua acurácia;
- Modificação do PSO para processamento paralelo; e
- Aplicação de grandes bases de dados com valores ausente.

REFERÊNCIAS

ABDELLA, M. e MARWALA, T. Treatment of missing data using neural networks and genetic algorithms. **Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.**, v. 2, p. 598–603, doi:10.1109/IJCNN.2005.1555899, 2005.

ABDELLA, M. e MARWALLA, T. The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data. **Computing and Informatics**, v. 24, p. 577–589, 2005.

ALCALÁ-FDEZ, J. et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems. **Soft Computing**, v. 13, n. 3, p. 307–318, doi:10.1007/s00500-008-0323-y, 2008.

ANDRADE SILVA, J. DE e HRUSCHKA, E. R. EACImpute: An Evolutionary Algorithm for Clustering-Based Imputation. **2009 Ninth International Conference on Intelligent Systems Design and Applications**, p. 1400–1406, doi:10.1109/ISDA.2009.86, 2009.

ASUNCION, A. e NEWMAN, D. J. **UCI Machine Learning Repository**. **University of California Irvine School of Information**. [S.l.]: University of California, Irvine, School of Information and Computer Sciences. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html>, 2007.

AZADEH, A. et al. Knowledge-Based Systems Optimum estimation of missing values in randomized complete block design by genetic algorithm. **Knowledge-Based Systems**, v. 37, p. 37–47, 2013.

BANSAL, J. C. et al. Inertia Weight Strategies in Particle Swarm. In: WORLD CONGRESS ON NATURE AND BIOLOGICALLY INSPIRED COMPUTING. **Anais**... [S.l: s.n.], 2011.

BATISTA, G. E. a. P. a. e MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5-6, p. 519–533, doi:10.1080/713827181, 2003.

BISHOP, C. M. Neural networks: a pattern recognition perspective. **Neural Networks**, n. 1973, p. 1–23, 1996.

BRÁS, L. P. e MENEZES, J. C. Improving cluster-based missing value estimation of DNA microarray data. **Biomolecular Engineering**, v. 24, n. 2, p. 273–282, 2007.

BROWN, M. L. e KROS, J. F. Data mining and the impact of missing data. **Industrial Management & Data Systems**, v. 103, n. 8, p. 611–621, doi:10.1108/02635570310497657, 2003.

CARACIOLO, M. **Artificial Intelligence in Motion: Introdução à Inteligência de Enxame - Otimização por Enxame de Partículas (PSO)**. Disponível em: <http://aimotion.blogspot.com.br/2009/04/introducao-inteligencia-de-enxame.html>. Acesso em: 20 maio. 2013.

CASTRO, L. N. DE. **Fundamentals of Natural Computing**. [S.l.]: Chapman and Hall, 2006. v. 4p. 662

CASTRO, L. N. DE. Fundamentals of natural computing: an overview. **Physics of Life Reviews**, v. 4, n. 1, p. 1–36, doi:10.1016/j.plrev.2006.10.002, 2007.

CHANG, G.;; ZHANG, Y. e YAO, D. Missing Data Imputation for Traffic Flow Based on Improved Local Least Squares *. **Tsinghua Science and Technology**, v. 17, n. 3, p. 304–309, 2012.

CHENG, K. O.;; LAW, N. F. e SIU, W. C. Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. **Pattern Recognition**, v. 45, n. 4, p. 1281–1289, doi:10.1016/j.patcog.2011.10.012, 2012.

DHLAMINI, S. M.;; NELWAMONDO, F. V e MARWALA, T. Condition Monitoring of HV Bushings in the Presence of Missing Data Using Evolutionary Computing. **CoRR**, p. 1–7, 2007.

DIAS, L. de J. C.;; LOBATO, F. M. F. e SANTANA, Á. L. DE. A Testbed for the Experiments Performed in Missing Value Treatments. **International Journal of Computer Science and Engineering**, v. 7, n. 8, p. 311–315, 2013.

DING, Y. e ROSS, A. A comparison of imputation methods for handling missing scores in biometric fusion. **Pattern Recognition**, v. 45, n. 3, p. 919–933, doi:10.1016/j.patcog.2011.08.002, 2012.

EBERHART, R. e KENNEDY, J. A new optimizer using particle swarm theory. In: INTERNATIONAL SYMPOSIUM ON MICRO MACHINE AND HUMAN SCIENCE. **Anais**... [S.l.]: Ieee, 1995.

EIBEN, A. E. e SMITH, J. E. **Introduction to Evolutionary Computing**. [S.l.]: Springer, 2003. v. 12p. 299

ENGELBRECHT, A. P. **Computational Intelligence: An Introduction**. [S.l: s.n.], 2007.

FACELI, K. et al. **Inteligência Artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LCT, 2011.

FARHANGFAR, A.;; KURGAN, L. a. e PEDRYCZ, W. Experimental analysis of methods for imputation of missing values in databases. **Intelligent Computing: Theory and Applications II**, v. 5421, p. 172–182, doi:10.1117/12.542509, 2004.

FARHANGFAR, A.;; KURGAN, L. a. e PEDRYCZ, W. A Novel Framework for Imputation of Missing Values in Databases. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, v. 37, n. 5, p. 692–709, doi:10.1109/TSMCA.2007.902631, 2007.

FIGUEROA GARCÍA, J. C.;; KALENATIC, D. e LOPEZ BELLO, C. A. Missing Data Imputation in Time Series by Evolutionary Algorithms. v. 5227, doi:10.1007/978-3-540-85984-0, 2008.

FIGUEROA GARCÍA, J. C.;; KALENATIC, D. e LOPEZ BELLO, C. A. Missing data imputation in multivariate data by evolutionary algorithms. **Computers in Human Behavior**, v. 27, n. 5, p. 1468–1474, doi:10.1016/j.chb.2010.06.026, 2011.

FIGUEROA GARCÍA, J. C.;; KALENATIC, D. e LÓPEZ BELLO, C. A. An Evolutionary Approach for Imputing Missing Data in Time Series. **Journal of Circuits, Systems and Computers**, v. 19, n. 01, p. 107–121, doi:10.1142/S0218126610006050, 2010.

FRANÇA, F. O. DE;; COELHO, G. P. e ZUBEN, F. J. VON. Predicting missing values with biclustering: A coherence-based approach. **Pattern Recognition**, v. 46, n. 5, p. 1255–1266, doi:10.1016/j.patcog.2012.10.022, 2013.

GONÇALVES DE OLIVEIRA, P. **Imputação automática de atributos faltantes em problemas de classificação: Um estudo comparativo envolvendo algoritmos bio-inspirados**. Universidade de Fortaleza - [S.l.]. 2009.

GRAHAM, J. W. Missing data analysis: making it work in the real world. **Annual review of psychology**, v. 60, p. 549–76, doi:10.1146/annurev.psych.58.110405.085530, 2009.

HALL, M. et al. The WEKA data mining software: an update. **SIGKDD Explorations**, v. 11, n. 1, p. 10–18, doi:10.1145/1656274.1656278, 2009.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. [S.l.]: Prentice Hall, 1999. v. 13p. 409–412

HEERINGA, S.;; WEST, B. e BERGLUND, P. **Applied survey data analysis**. [S.l.]: Chapman & Hall/CRC, 2010.

HONGHAI, F.;; GUOSHUN, C. e CHENG, Y. A SVM Regression Based Approach to Filling in Missing Values. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED AND INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS. **Anais**... [S.l: s.n.], 2005.

KENNEDY, J. e EBERHART, R. Particle swarm optimization. **IEEE International Conference on Neural Networks**, v. 1, n. 1, p. 1942–1948, doi:10.1007/s11721-007-0002-0, 1995.

KENNEDY, J.;; EBERHART, R. e SHI, Y. **Swarm intelligence**. [S.l: s.n.], 2001. v. 78

KIM, K.-Y.;; KIM, B.-J. e YI, G.-S. Reuse of imputed data in microarray analysis increases imputation efficiency. **BMC Bioinformatics**, v. 5, n. 1, p. 160, 2004.

LAKSHMINARAYAN, K.;; HARP, S. A. e SAMAD, T. Imputation of Missing Data in Industrial Databases. **Applied Intelligence**, v. 275, p. 259–275, 1999.

LI, X. e WANG, Q. The weighted least square based estimators with censoring indicators missing at random. **Journal of Statistical Planning and Inference**, v. 142, n. 11, p. 2913–2925, doi:10.1016/j.jspi.2012.04.016, 2012.

LITTLE, R. J. A. e RUBIN, D. B. **Statistical Analysis with Missing Data**. 2. ed. New York: Wiley, 2002. v. Secondp. 408

LIU, Y. e BROWN, S. D. Comparison of five iterative imputation methods for multivariate classification. **Chemometrics and Intelligent Laboratory Systems**, v. 120, p. 106–115, 2013.

LUENGO, J.;; SÁEZ, J. a. e HERRERA, F. Missing data imputation for fuzzy rule-based classification systems. **Soft Computing**, v. 16, n. 5, p. 863–881, doi:10.1007/s00500-011-0774-4, 2012.

MANGE, D. e TOMASSINI, M. **Bio-inspired computing machines**. [S.l: s.n.], 1998.

MCCULLOCH, W. S. e PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, doi:10.1007/BF02478259, 1943.

MCKNIGHT, P. E. et al. **Missing Data: A Gentle Introduction**. [S.l: s.n.], 2007.

PATIL, D. V. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques. In: INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS. **Anais**... [S.l: s.n.], 2010.

PATON, R. **Computing with biological metaphors**. [S.l.]: Chapman & Hall, 1994. p. 446

POLI, R. Analysis of the Publications on the Applications of Particle Swarm Optimisation. **Journal of Artificial Evolution and Applications**, n. 2, p. 1–10, doi:10.1155/2008/685175, 2008.

RUBIN, B. D. Inference and missing data. **Biometrika**, v. 63, n. 3, p. 581–592, 1976.

SCHAFER, J. L. **Analysis of Incomplete Multivariate Data**. [S.l.]: Chapman & Hall, 1997. v. 11p. 164–165

SCHAFER, J. L. e GRAHAM, J. W. Missing data: Our View of the State of the Art. **Psycological Methods**, v. 7, n. 2, p. 147–177, 2002.

SHI, Y. e EBERHART, R. A modified particle swarm optimizer. In: IEEE INTERNATIONAL CONFERENCE ON EVOLUTIONARY COMPUTATION PROCEEDINGS. IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE. **Anais**... [S.l.]: Ieee, 1998.

SILVA, J. D. A. **Subistituição de valores ausentes: uma abordagem baseada em um algoritmo evolutivo para agrupamento de dados**. Instituto de Ciências Matemáticas e de Computação - Univerisdade de São Paulo - [S.l.]. 2010.

SILVA, J. D. A. e HRUSCHKA, E. R. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. **Data & Knowledge Engineering**, v. 84, p. 47–58, 2013.

SILVA-RAMÍREZ, E.-L. et al. Missing value imputation on missing completely at random data using multilayer perceptrons. **Neural networks**, v. 24, n. 1, p. 121–9, doi:10.1016/j.neunet.2010.09.008, 2011.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. **Bioinformatics**, v. 17, n. 6, p. 520–5, doi:10.1093/bioinformatics/17.6.520, 2001.

VERONEZE, R. et al. Assessing the Performance of a Swarm-based Biclustering Technique for Data Imputation. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION. **Anais**... [S.l: s.n.], 2011.

VERONEZE, R. **Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla**. Universidade Estadual de Campinas - [S.l.]. 2011.

WAYMAN, J. C. Multiple Imputation For Missing Data : What Is It And How Can I Use It ? In: ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. **Anais**... [S.l: s.n.], 2003.

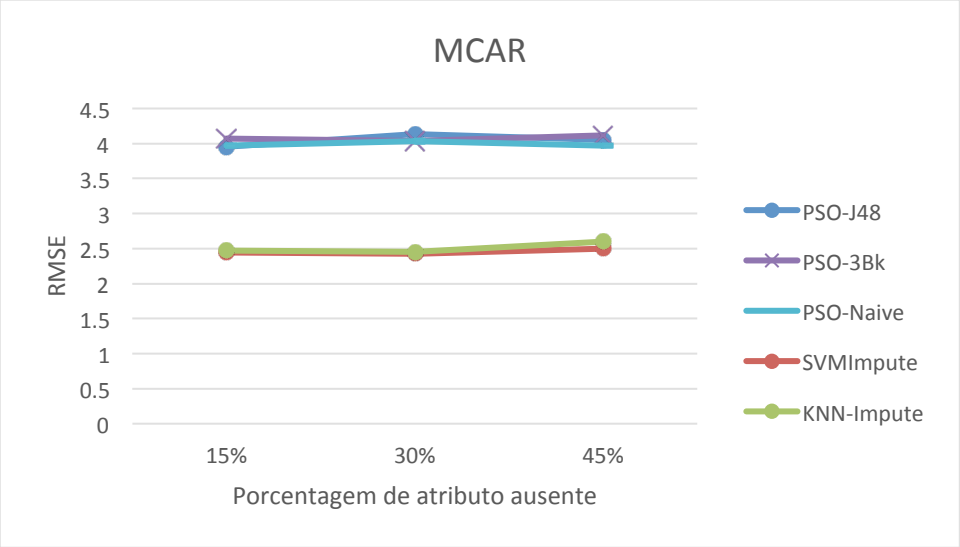
WOHLRAB, L. e FÜRNKRANZ, J. A review and comparison of strategies for handling missing values in separate-and-conquer rule learning. **Journal of Intelligent Information Systems**, v. 36, n. 1, p. 73–98, doi:10.1007/s10844-010-0121-8, 2010.

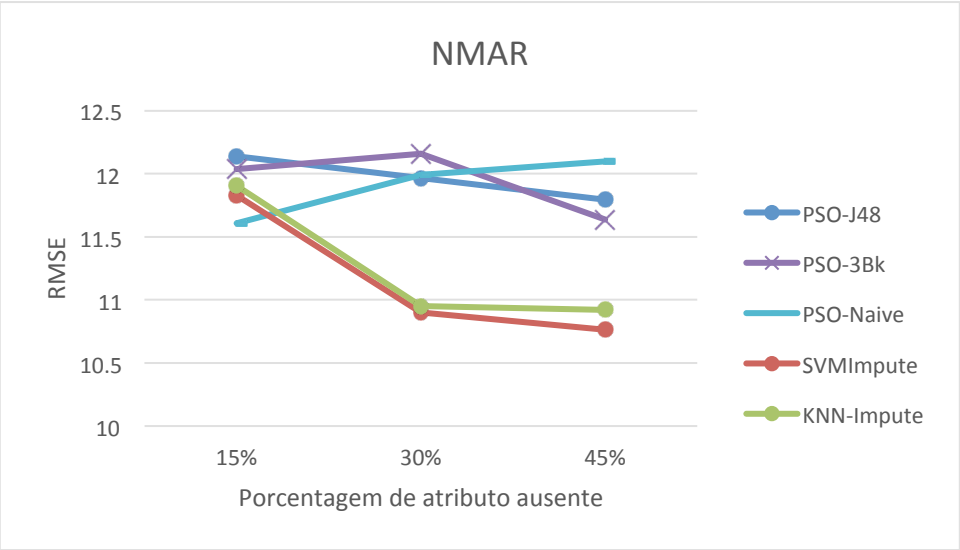
ZHANG, S. Nearest neighbor selection for iteratively kNN imputation. **Journal of Systems and Software**, v. 85, n. 11, p. 2541–2552, doi:10.1016/j.jss.2012.05.073, 2012.

APÊNDICE A – LISTA DOS RESULTADOS DOS EXPERIMENTOS POR BASE

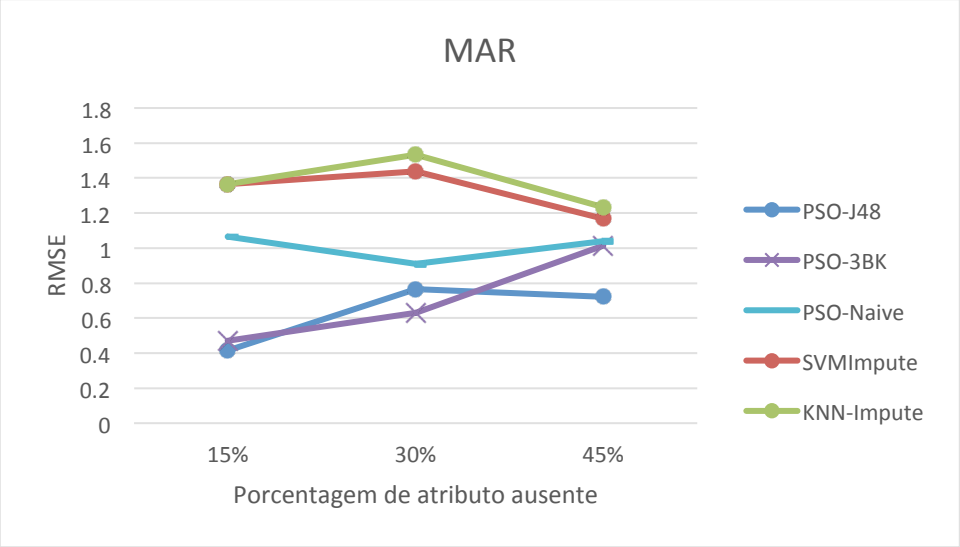
Neste documento estão listados os resultados dos experimentos realizados com os métodos de imputação de dados ausentes citados nesta dissertação, cada gráfico é apresentado para cada base de dados, as quais estão alfabeticamente ordenadas.

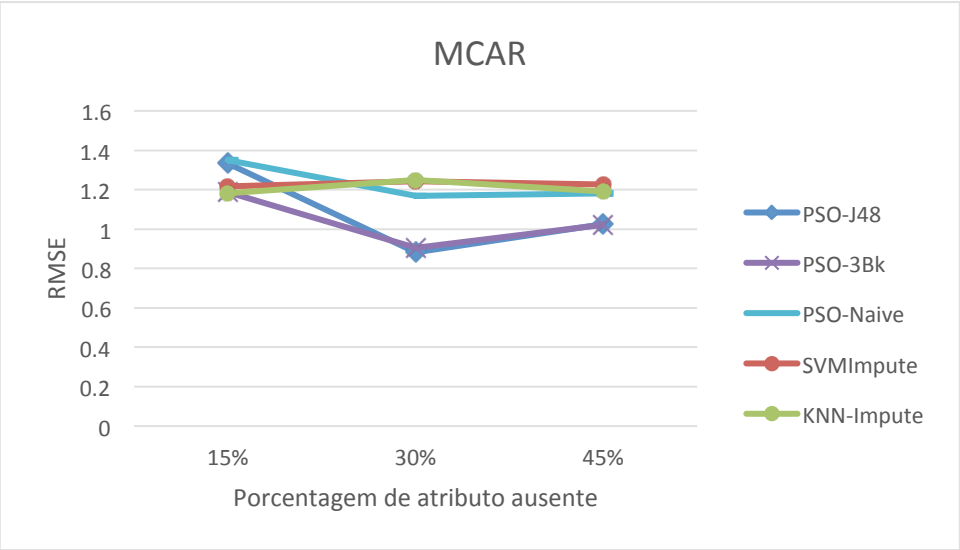
- Contraceptive

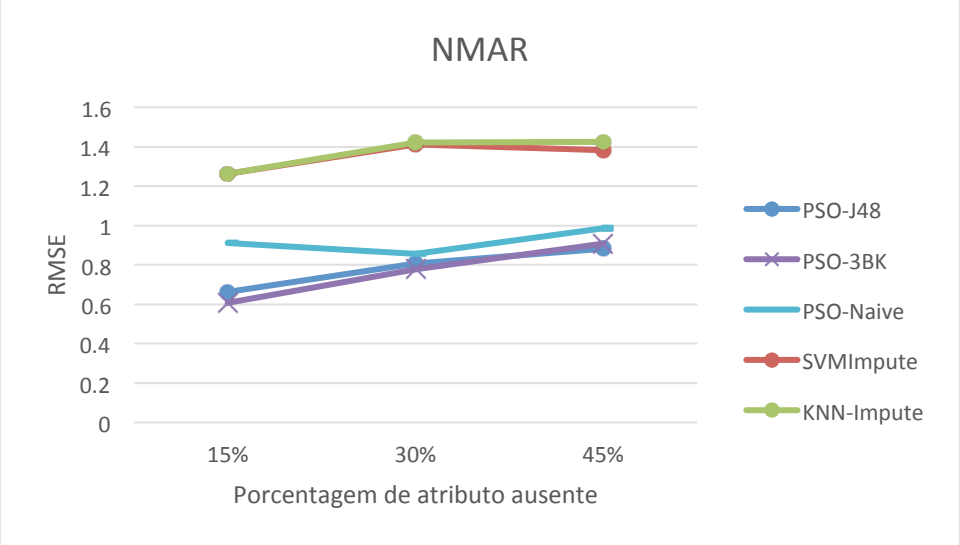



NMAR
12.5
12
11.5
11
10.5
10
RMSE
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3Bk
PSO-Naive
SVMImpute
KNN-Impute

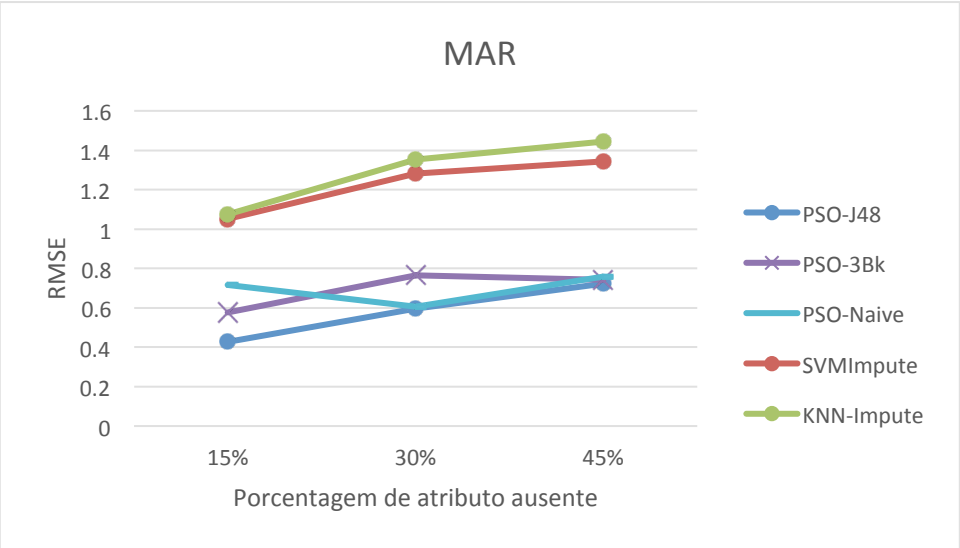
- Glass

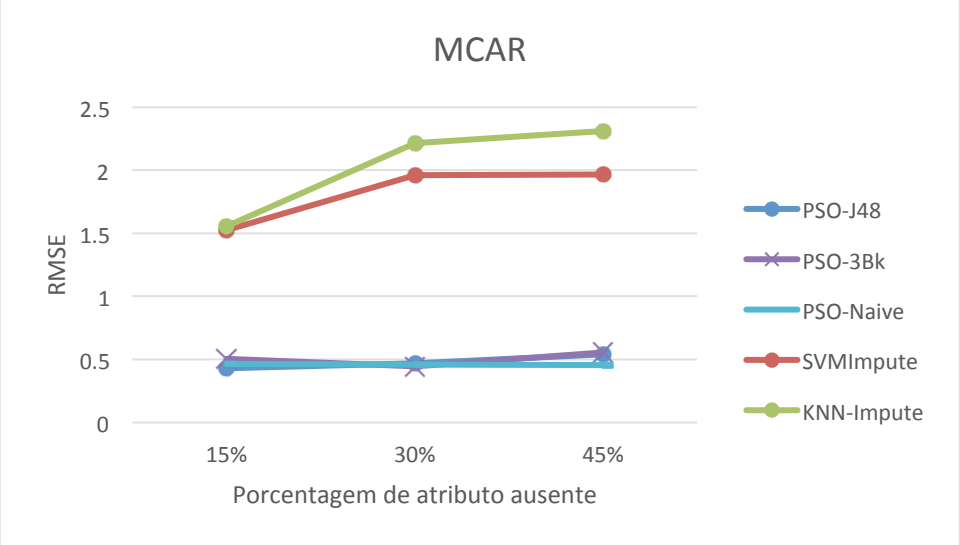

MAR
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
RMSE
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute


MCAR
RMSE
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3Bk
PSO-Naive
SVMImpute
KNN-Impute

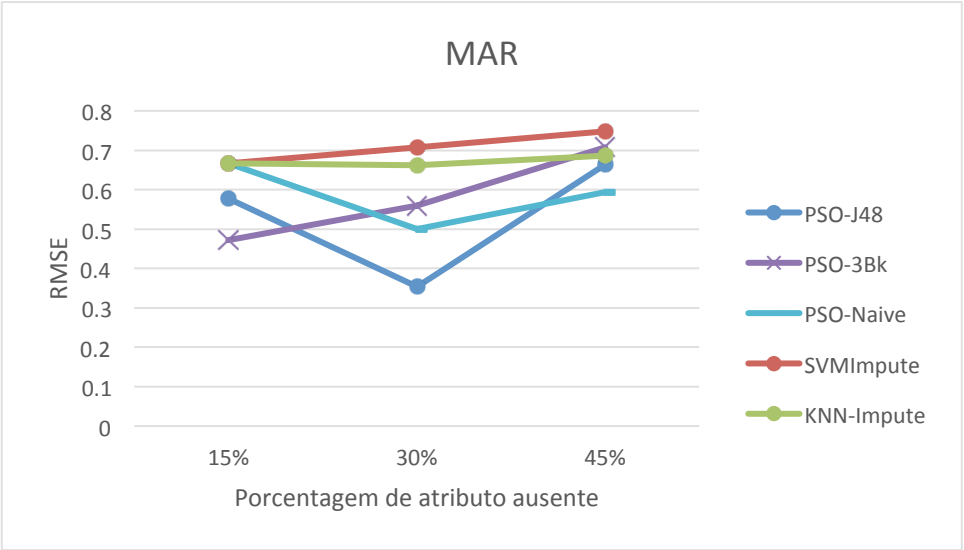

NMAR
RMSE
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute

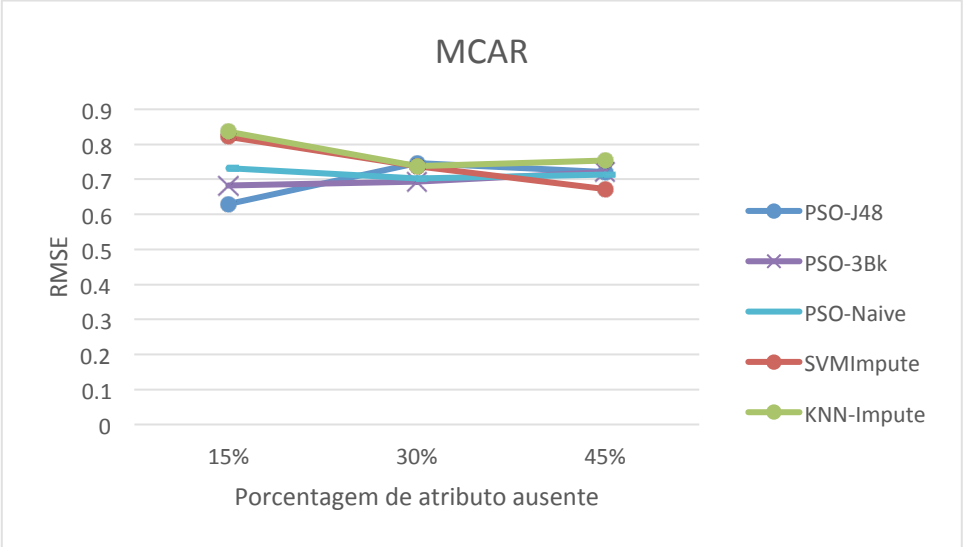
- Íris

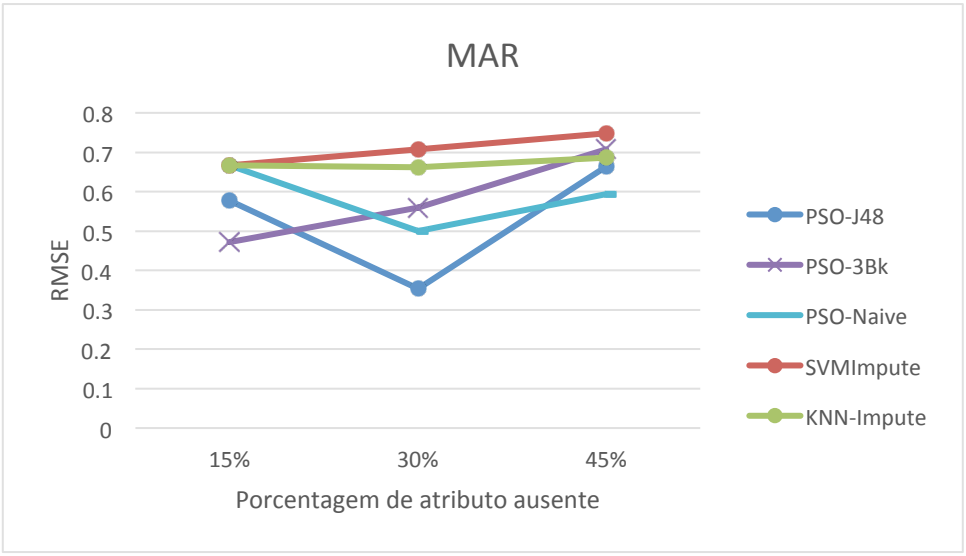




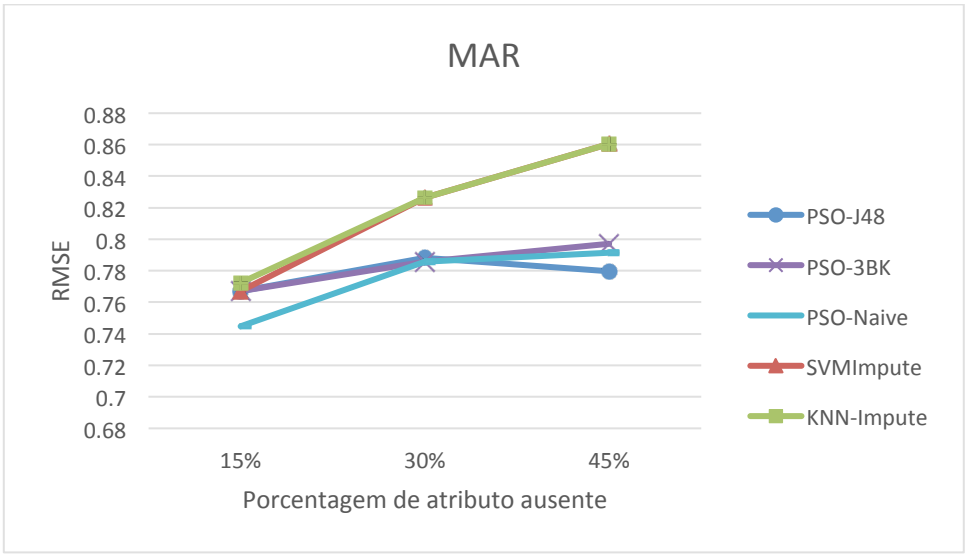
- Lymphography

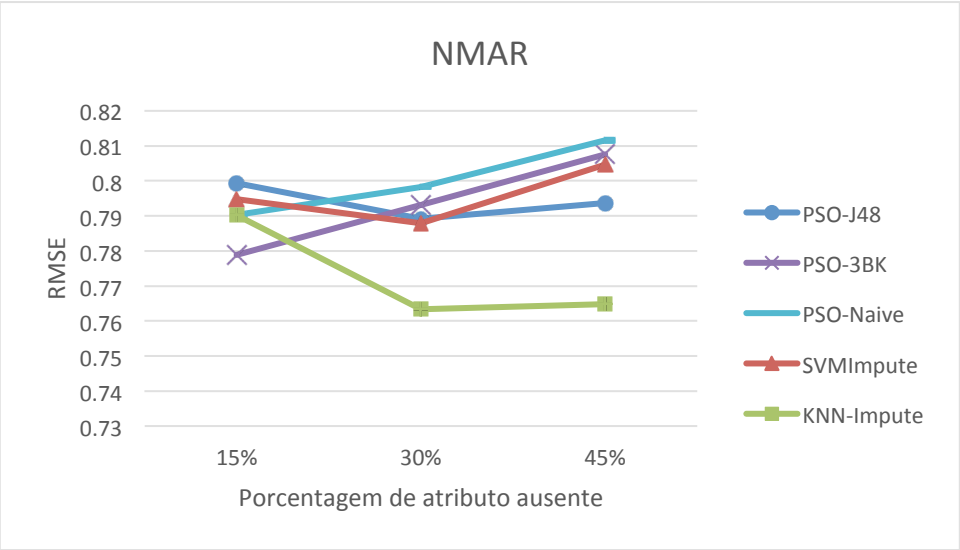

MAR
RMSE
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3Bk
PSO-Naive
SVMImpute
KNN-Impute


MCAR
RMSE
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3Bk
PSO-Naive
SVMImpute
KNN-Impute

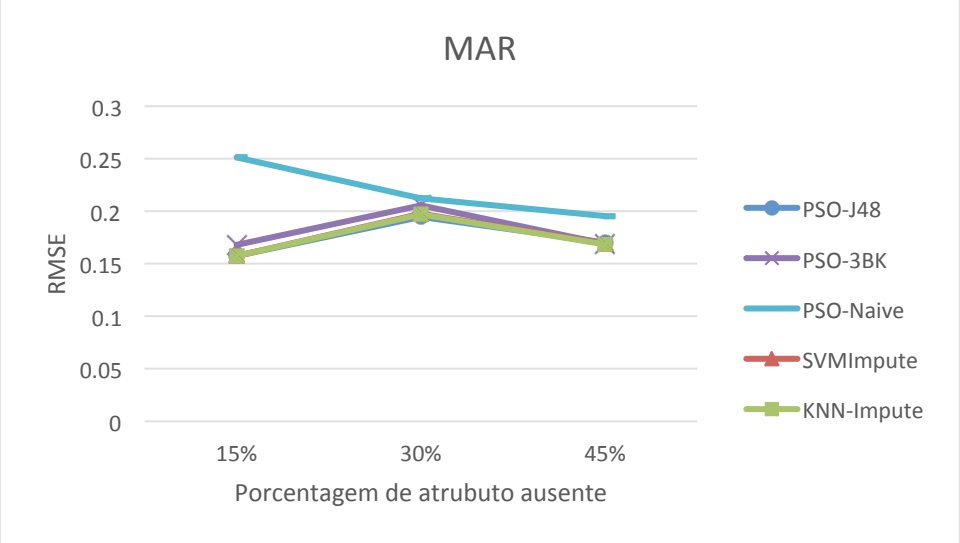

MAR
RMSE
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3Bk
PSO-Naive
SVMImpute
KNN-Impute

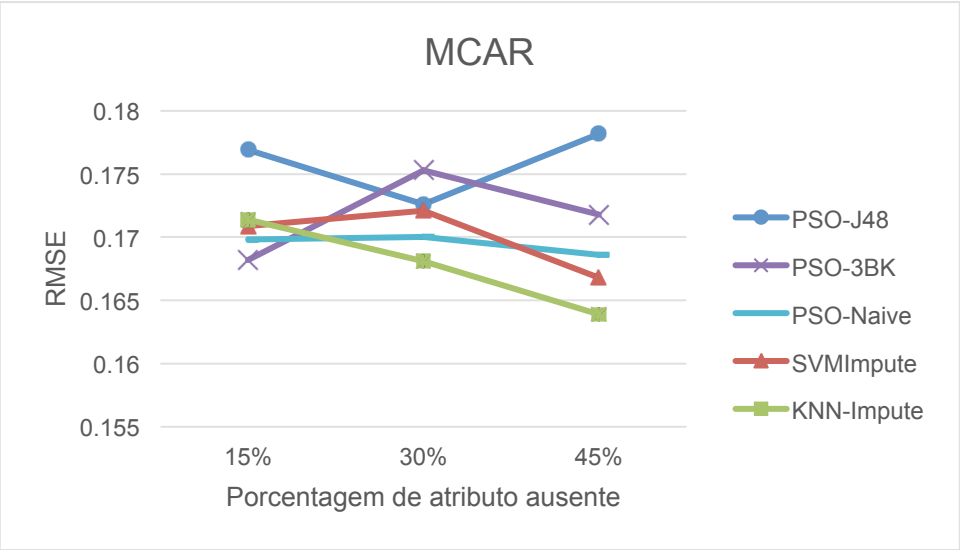
- Tic-tac-toe

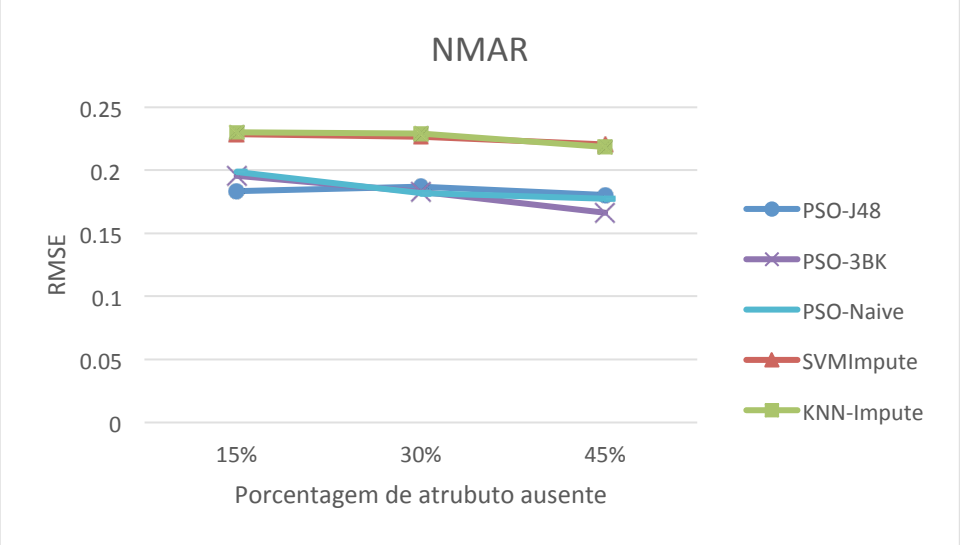

MAR
RMSE
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute


NMAR
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
0.74
0.73
RMSE
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute

- Yeast


MAR
0.3
0.25
0.2
0.15
0.1
0.05
0
RMSE
15%
30%
45%
Porcentagem de atrubuto ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute


MCAR
0.18
0.175
0.17
0.165
0.16
0.155
RMSE
15%
30%
45%
Porcentagem de atributo ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute


NMAR
0.25
0.2
0.15
0.1
0.05
0
RMSE
15%
30%
45%
Porcentagem de atrubuto ausente
PSO-J48
PSO-3BK
PSO-Naive
SVMImpute
KNN-Impute

APÊNDICE B – CARACTERÍSTICAS DAS BASES DE DADOS COM VALORES AUSENTES CONSIDERANDO OS MECANISMOS DE AUSÊNCIA

Este apêndice apresenta o quadro mostra a porcentagem dos atributos ausentes e, como consequência, a porcentagem de valores ausentes, e a de instâncias com VA nas bases geradas.

Nome	%VA	%Inst. VA
Glass_MCAR_15	2,99	27,10
Glass_MCAR_30	5,98	51,40
Glass_MCAR_45	8,97	68,69
Glass_MAR_15	0,32	3,27
Glass_MAR_30	0,56	5,60
Glass_MAR_45	1,26	12,61
Glass_NMAR_15	0,74	7,47
Glass_NMAR_30	1,54	15,42
Glass_NMAR_45	2,10	21,02
Lymphographic_MCAR_15	1,56	29,05
Lymphographic_MCAR_30	3,12	53,37
Lymphographic_MCAR_45	4,69	70,27
Lymphographic_MAR_15	0,64	12,16
Lymphographic_MAR_30	0,56	10,81
Lymphographic_MAR_45	1,2	22,97

Lymphographic_NMAR_15	0,96	18,24
Lymphographic_NMAR_30	2,09	39,86
Lymphographic_NMAR_45	1,92	36,48
Contraceptive_MCAR_15	2,98	27,76
Contraceptive_MCAR_30	5,98	51,18
Contraceptive_MCAR_45	8,98	69,78
Contraceptive_MAR_15	0,57	5,70
Contraceptive_MAR_30	1,14	11,47
Contraceptive_MAR_45	1,73	17,37
Contraceptive_NMAR_15	1,80	18,05
Contraceptive_NMAR_30	3,57	35,77
Contraceptive_NMAR_45	5,09	50,91
Iris_MCAR_15	5,86	28,66
Iris_MCAR_30	12	52,66
Iris_MCAR_45	17,86	69,66
Iris_MAR_15	0,93	4,66
Iris_MAR_30	1,73	8,66
Iris_MAR_45	2,13	10,66

Iris_NMAR_15	3,33	16,66
Iris_NMAR_30	4,53	22,66
Iris_NMAR_45	7,33	36,66
Tic-tac-toe_MCAR_15	4,47	39,56
Tic-tac-toe_MCAR_30	8,98	63,88
Tic-tac-toe_MCAR_45	13,49	84,02
Tic-tac-toe_MAR_15	1,38	12,42
Tic-tac-toe_MAR_30	2,96	25,36
Tic-tac-toe_MAR_45	4,40	33,82
Tic-tac-toe_NMAR_15	3,11	28,91
Tic-tac-toe_NMAR_30	5,87	49,26
Tic-tac-toe_NMAR_45	8,31	65,13
Yeast_MCAR_15	4,48	38,07
Yeast_MCAR_30	8,99	64,48
Yeast_MCAR_45	13,48	82,34
Yeast_MAR_15	0,17	1,75
Yeast_MAR_30	0,42	4,24
Yeast_MAR_45	0,72	7,21
Yeast_NMAR_15	0,51	5,05
Yeast_NMAR_30	0,82	8,08
Yeast_NMAR_45	1,39	13,94

Vertebral Column_MCAR_15	4,23	27,74
Vertebral Column_MCAR_30	8,57	51,29
Vertebral Column_MCAR_45	12,81	70,32
Vertebral Column_MAR_15	2,11	14,83
Vertebral Column_MAR_30	2,99	20,96
Vertebral Column_MAR_45	6,49	45,48
Vertebral Column_NMAR_15	2,02	14,19
Vertebral Column_NMAR_30	3,64	25,48
Vertebral Column_NMAR_45	5,76	40,32