FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MESTRADO EM INFORMÁTICA APLICADA

Pedro Gonçalves de Oliveira

IMPUTAÇÃO AUTOMÁTICA DE ATRIBUTOS FALTANTES EM PROBLEMAS DE CLASSIFICAÇÃO: UM ESTUDO COMPARATIVO ENVOLVENDO ALGORITMOS BIO-INSPIRADOS

Fortaleza
2009

FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MESTRADO EM INFORMÁTICA APLICADA

Pedro Gonçalves de Oliveira

IMPUTAÇÃO AUTOMÁTICA DE ATRIBUTOS FALTANTES EM PROBLEMAS DE CLASSIFICAÇÃO: UM ESTUDO COMPARATIVO ENVOLVENDO ALGORITMOS BIO-INSPIRADOS

Dissertação apresentada ao Curso de Mestrado em Informática Aplicada da Universidade de Fortaleza como requisito parcial para obtenção do Título de Mestre em Informática.

Orientador: Prof. Dr. André Luís Vasconcelos Coelho

Fortaleza
2009

Pedro Gonçalves de Oliveira

IMPUTAÇÃO AUTOMÁTICA DE ATRIBUTOS FALTANTES EM PROBLEMAS DE CLASSIFICAÇÃO: UM ESTUDO COMPARATIVO ENVOLVENDO ALGORITMOS BIO-INSPIRADOS

Data de Aprovação __/__/____

Banca Examinadora:

Prof. André Luís Vasconcelos Coelho, D. Sc.
(Orientador: Universidade de Fortaleza – Unifor)

Prof. André Ponce de Leon F. de Carvalho, Ph.D.
(Membro Externo: Universidade de São Paulo – USP/SC)

Prof. Plácido Rogério Pinheiro, D. Sc.
(Membro Interno: Universidade de Fortaleza – Unifor)

Aos meus amados pais
e à minha querida esposa.

AGRADECIMENTOS

Aos meus pais, Pedro e Graça, e meu irmão Daniel, por tudo que ensinaram, pelo apoio que sempre me deram e pelos ótimos momentos em família. São meu porto seguro, meus pilares.

A minha esposa e amiga Thais, por estar sempre ao meu lado me dando a força necessária para seguir em frente e me ajudando a superar os desafios que aparecem na minha vida. Sem ela nada seria possível.

Ao professor André Luís Vasconcelos Coelho, pela orientação, ensinamentos, paciência e confiança em mim depositada. Por ter me motivado durante o curso com sua competência e visível admiração pela pesquisa. Pelas críticas, sempre construtivas, fundamentais para a minha formação.

A todos os colegas e professores da Unifor, que ajudaram em minha formação acadêmica. Aos amigos que constituem a célula Engenharia do Conhecimento, pela convivência e apoio constante.

A todos do NATI e da DATAPREV, pessoas fundamentais para realização deste trabalho.

A todos meus amigos que, de forma direta ou indireta, tornaram possível a realização deste trabalho. Amigos com os quais compartilhei conhecimento, alegrias e tristezas. Que me dão o incentivo necessário para completar qualquer desafio e me enchem de alegria pelo simples motivo de existirem.

Resumo da dissertação apresentada ao Corpo Docente do programa de Mestrado em Informática Aplicada da Universidade de Fortaleza, como parte dos requisitos necessários para a obtenção do grau de Mestre em Informática.

Imputação automática de atributos faltantes em problemas de classificação: Um estudo comparativo envolvendo algoritmos bio-inspirados

Autor: Pedro Gonçalves de Oliveira
Orientador: Prof. Dr. André Luís Vasconcelos Coelho

Diversas bases de dados reais se caracterizam pela ausência marcante de determinados valores de seus atributos. Esses dados ausentes podem vir a degradar sobremaneira o desempenho de algoritmos de mineração de dados, dificultando a análise resultante. Uma maneira comum de tratar esse problema é via imputação, ou seja, estimação dos valores faltantes a partir de outros existentes na base. Este trabalho avalia como uma abordagem de imputação por otimização numérica utilizando algoritmos bio-inspirados pode vir a aprimorar o desempenho de classificadores induzidos sobre as bases pré-processadas. Três técnicas foram empregadas segundo esta abordagem: imputação utilizando algoritmo genético (GA), imputação utilizando otimização por enxame de partículas (PSO) e imputação utilizando co-evolução cooperativa. Com o intuito de analisar as técnicas propostas, em termos de eficiência e eficácia, seis bases de dados do repositório UCI e cinco populares algoritmos de classificação foram adotados. Para efeito de comparação, foram empregadas ainda outras duas técnicas tradicionais de imputação: a imputação pela média ou moda e a imputação fazendo uso do algoritmo KNN. O estudo mostra que todas as técnicas de imputação consideradas são capazes de elevar o desempenho dos classificadores. Os resultados obtidos não apontam para um método ótimo para todas as situações. Contudo, os experimentos sugerem que, em geral, as técnicas que fazem uso de algoritmos bio-inspirados são as mais eficazes ao passo que as técnicas tradicionais são as de melhor desempenho computacional. Observa-se também que os algoritmos co-evolução cooperativa e PSO, ainda não muito explorados no contexto de pré-processamento de dados, sobressaíram-se em diversos experimentos realizados.

Palavras-Chave: Mineração de dados, Pré-processamento de dados, Imputação de valores faltantes, Metaheurísticas, Algoritmos genéticos, Otimização por enxame de partículas, Co-evolução cooperativa.

Abstract of the dissertation presented to the board of faculties of the Master Program in Applied Informatics at the University of Fortaleza, as partial fulfillment of the requirements for the Master's degree in Computer Science.

Automatic missing value imputation in classification problems: A comparative study using bio-inspired algorithms

Author: Pedro Gonçalves de Oliveira
Advisor: Prof. Dr. André Luís Vasconcelos Coelho

Real-world databases may contain several missing values, which may degrade the performance of data mining algorithms running over them, making it hard to analyze data. This problem is usually dealt with missing value imputation. The present work evaluates how imputation by numerical optimization using bio-inspired algorithms may affect the performance of classifiers induced over pre-processed data. Here, three techniques were conceived: imputation using genetic algorithm (GA), imputation using particle swarm optimization (PSO), and imputation using cooperative co-evolution. In order to analyze the proposed techniques, six different datasets from the UCI Machine Learning Repository and five well known classification algorithms were adopted. In this analysis, efficiency and efficacy criteria were taken into account. In order to compare the results obtained, two traditional missing value imputation techniques were used, namely, imputation using mean or mode, and imputation using the KNN algorithm. The study shows that all the imputation techniques considered could increase the performance of the resulting classifiers. The obtained results do not point out an optimal method, adequate to all situations. The experiments, however, showed that, in general, the techniques that use bio-inspired algorithms were the most effective, while traditional techniques entailed better computational performance. It should also be observed that the heuristic techniques PSO and cooperative co-evolution, still not much explored in the context of data preprocessing, could have prevailed in several experiments.

Keywords: Data mining, Data preparation, Missing value imputation, Metaheuristics, Genetic algorithms, Particle swarm optimization, Cooperative co-evolution.

SUMÁRIO

LISTA DE FIGURAS 10
LISTA DE TABELAS 11
1. INTRODUÇÃO 12
2. PRÉ-PROCESSAMENTO DE DADOS PARA MINERAÇÃO 15
2.1. Limpeza dos dados 17
2.1.1. Valores faltantes 17
2.1.1.1. Imputação de valores faltantes 20
2.1.2. Ruídos 21
2.2. Transformação dos dados 23
2.3. Redução dos dados 23
2.3.1. Seleção de atributos 24
2.3.2. Seleção de instâncias 25
2.4. Conclusão 26
3. ALGORITMOS HEURÍSTICOS BIO-INSPIRADOS 27
3.1. Algoritmos genéticos 28
3.2. Otimização por enxame de partículas 31
3.3. Co-evolução cooperativa 34
3.4. Conclusão 37
4. EXPERIMENTOS COMPUTACIONAIS 38
4.1. Técnicas de imputação adotadas 42
4.1.1. Imputação utilizando Algoritmos Genéticos 45
4.1.2. Imputação utilizando PSO 46
4.1.3. Imputação utilizando co-evolução 47
4.2. Resultados 48
4.2.1. Resultados para redes neurais RBF 48
4.2.2. Resultados para o algoritmo J48 53

4.2.3. Resultados para o algoritmo IBK 57
4.2.4. Resultados para o algoritmo Naïve Bayes 61
4.2.5. Resultados para o algoritmo PART 65
4.3. Conclusão 69
5. CONSIDERAÇÕES FINAIS 70
BIBLIOGRAFIA 72
Apêndice I – Resultados dos experimentos computacionais 78

LISTA DE FIGURAS

Figura 1: Esquema geral KDD (Fayyad, Gregory e Padhraic 1996) 16
Figura 2: Agrupamentos com ruídos 22
Figura 3: Esquema básico de um algoritmo genético típico 30
Figura 4: Arquitetura de co-evolução 35
Figura 5: Geração das bases para o experimento 39
Figura 6: Exemplo de árvore de decisão 40
Figura 7: Exemplo de rede neural RBF 41
Figura 8: Representação do indivíduo para algoritmo genético 46
Figura 9: Representação da partícula para PSO 46
Figura 10: Representação do indivíduo para co-evolução 47
Figura 11: Resultados obtidos com redes neurais RBF – Iris, Zoo e Diabetes (validação cruzada) 50
Figura 12: Resultados obtidos com redes neurais RBF – BCD699, Car e Dermatology (validação cruzada) 51
Figura 13: Resultados obtidos para o algoritmo J48 – Iris, Zoo e Diabetes (validação cruzada) 54
Figura 14: Resultados obtidos para o algoritmo J48 – BCD699, Car e Dermatology (validação cruzada) 55
Figura 15: Resultados obtidos com o algoritmo IBK – Iris, Zoo e Diabetes (validação cruzada) 58
Figura 16: Resultados obtidos com o algoritmo IBK – BCD699, Car e Dermatology (validação cruzada) 59
Figura 17: Resultados obtidos com o algoritmo Naïve Bayes – Iris, Zoo e Diabetes (validação cruzada) 62
Figura 18: Resultados obtidos com o algoritmo Naïve Bayes – BCD699, Car e Dermatology (validação cruzada) 63
Figura 19: Resultados obtidos com o algoritmo PART – Iris, Zoo e Diabetes (validação cruzada) 66
Figura 20: Resultados obtidos com o algoritmo PART – BCD699, Car e Dermatology (validação cruzada) 67

LISTA DE TABELAS

Tabela 1: Pseudocódigo de um algoritmo genético típico 29
Tabela 2: Pseudocódigo do PSO padrão 33
Tabela 3: Bases de dados utilizadas nos experimentos 38
Tabela 4: Tabela comparativa de trabalhos experimentais relacionados 44
Tabela 5: Tempo de pré-processamento para redes neurais RBF 52
Tabela 6: Tempo de pré-processamento para o algoritmo J48 56
Tabela 7: Tempo de pré-processamento para o algoritmo IBK 60
Tabela 8: Tempo de pré-processamento para o algoritmo Naïve Bayes 64
Tabela 9: Tempo de pré-processamento (em segundos) para o algoritmo PART 68

1. INTRODUÇÃO

Grandes volumes de dados estão cada vez mais presentes em nossas vidas. Muitas vezes se deseja extrair dessas massas de dados informações que estão intrínsecas e que podem ser potencialmente úteis. Para realizar essa extração de conhecimento de forma automática, mais recentemente vem sendo comum a adoção em larga escala de algoritmos de mineração de dados (Han e Kamber 2001) (Witten e Frank 2005).

No entanto, muitas dessas bases de dados possuem problemas que podem vir a degradar sobremaneira o desempenho dos algoritmos de mineração. Para minimizar esses problemas, diversos métodos de pré-processamento de dados vêm sendo empregados (Pyle 1999) (Zhang, Yang e Zhang 2002). Considera-se a qualidade dos dados a serem minerados um fator primordial para se obter resultados relevantes. Nesse contexto, o presente trabalho foca especificamente no problema de tratamento de valores faltantes, o qual vem sendo recorrente em diferentes configurações de bases de dados reais (Farhangfar, Kurgan e Dy 2008).

Os valores faltantes podem ocorrer por diversos motivos: erros no preenchimento, falhas na digitação, problemas na digitalização de documentos, etc. Existem diversas abordagens capazes de tratar esse problema (Little e Rubin 1987) (Schafer e Graham 2002). Todavia, este trabalho gira em torno exclusivamente de abordagens de imputação, mediante as quais os valores faltantes passam a ser estimados, tomando como referencial, geralmente, os demais valores presentes na base de dados. Diversas técnicas de imputação vêm sendo propostas na literatura (Farhangfar, Kurgan e Dy 2008). O trabalho aqui apresentado tem como objetivo investigar técnicas baseadas em algoritmos bio-inspirados, comparando-os com técnicas de imputação tradicionais, particularmente a substituição pela média ou moda e a substituição empregando o algoritmo KNN (Batista e Monard 2003).

A escassez de trabalhos a respeito de imputação de valores faltantes no contexto dos algoritmos bio-inspirados (de Castro 2007) e os bons resultados alcançados por algoritmos evolutivos em diversas tarefas de pré-processamento de dados (Freitas 2002) foram as maiores motivações por trás deste trabalho, o qual propõe a modelagem da tarefa de imputação de valores faltantes como um problema típico de otimização numérica[1]. Essa abordagem de imputação baseada em otimização, por outro lado, é instanciada mediante o emprego de três algoritmos heurísticos bio-inspirados, dando origem a três técnicas de imputação diferentes.

A primeira técnica de imputação bio-inspirada, empregando um algoritmo genético (GA) (Holland 1975), usa conceitos inspirados na teoria da evolução das espécies e em genética populacional, tais como os de seleção natural, reprodução, mutação, e hereditariedade. Foram usadas três diferentes configurações de GA, uma tradicional e duas novas propostas neste trabalho. As implementações propostas aqui alteram apenas o operador de mutação. A segunda técnica de imputação bio-inspirada, empregando otimização por enxame de partículas (PSO) (Kennedy e Eberhart 1995), baseia-se na simulação de comportamentos coletivos exibidos por espécies sociais, tais como cardumes de peixes, revoadas de pássaros e matilhas de lobos. A terceira técnica de imputação, que faz uso do conceito de co-evolução cooperativa (Potter e De Jong 2000) (Coelho 2004), trabalha com dois ou mais algoritmos genéticos simultaneamente, cada um representando uma espécie no ecossistema e responsável por imputar parte dos valores faltantes. Essas espécies evoluem simbioticamente de forma a gerar uma base de dados final única, segundo o princípio da "divisão-e-conquista".

Como forma de avaliar as novas técnicas bio-inspiradas de imputação, foram realizados testes com bases de dados populares provenientes do repositório UCI (Asuncion e Newman 2007). Essas bases foram manipuladas, de maneira controlada, a fim de se introduzir valores faltantes e avaliar, por conseguinte, o desempenho de classificadores sobre as bases tratadas pelos métodos de imputação. Para cada base, foram geradas variações contendo 5%, 10%, 20% e 33% de valores faltantes, de forma a possibilitar a apuração do impacto da quantidade de valores faltantes na eficiência e

[1] Uma vez que algoritmos genéticos e outros métodos de cunho heurístico são também considerados como métodos de Aprendizado de Máquina (Mitchell, 1997), essa modelagem também poderia ser realizada sob esse prisma, ficando a abordagem de imputação bio-inspirada resultante próxima àquela baseada no algoritmo KNN (Batista e Monard 2003). Todavia, neste documento, preferiu-se considerar a tarefa de imputação sob a ótica de otimização numérica.

eficácia dos algoritmos. Foram empregados nos experimentos cinco classificadores amplamente adotados na literatura, cada qual representando um tipo diferente de indutor (Witten e Frank 2005): um algoritmo baseado em árvores de decisão, o J48; um algoritmo baseado na estatística bayesiana, o Naïve Bayes; um algoritmo baseado em regras de decisão, o PART; um algoritmo baseado em funções numéricas, i.e., redes neurais do tipo base radial (RBF); e, por fim, um algoritmo de aprendizado "preguiçoso" baseado em instâncias, o IBK.

Os experimentos apontam que, em geral, a abordagem de imputação por otimização numérica utilizando heurísticas bio-inspiradas é a mais eficaz, no sentido de acarretar maiores ganhos de acuidade (precisão) por parte dos classificadores induzidos. Entretanto, os métodos bio-inspirados de imputação são os que apresentam pior desempenho computacional. Os resultados obtidos não apontam para uma técnica ótima para todas as situações. As técnicas empregando PSO e co-evolução cooperativa se sobressaíram por diversas vezes, encontrando a melhor solução de imputação para os diferentes pares de classificadores e bases de dados. As técnicas tradicionais aplicadas dificilmente predominaram sobre as bio-inspiradas; contudo, apresentaram um melhor desempenho computacional.

O trabalho se organiza da seguinte forma. No Capítulo 2, são apresentados alguns conceitos a respeito de mineração de dados bem como as tarefas de pré-processamento de dados, dando maior ênfase à tarefa de tratamento de valores faltantes. No Capítulo 3, são descritos os algoritmos heurísticos bio-inspirados empregados em nossas investigações: algoritmos genéticos, PSO e co-evolução cooperativa. No Capítulo 4, são apresentados os experimentos conduzidos neste trabalho, assim como é feita uma análise detalhada dos resultados obtidos. O Capítulo 5 é destinado à conclusão e à discussão de trabalhos futuros.

2. PRÉ-PROCESSAMENTO DE DADOS PARA MINERAÇÃO

Neste capítulo, serão apresentados alguns conceitos relacionados ao processo de descoberta automática de conhecimento em bases de dados, dando maior ênfase às tarefas de pré-processamento de dados, notadamente à de tratamento de valores faltantes. O pré-processamento faz parte do processo de descoberta de conhecimento em base de dados, também conhecido como KDD[2] (Han e Kamber 2001) (Witten e Frank 2005).

O KDD pode ser dividido em diversas etapas, como pode ser visto na Figura 1. O processo se inicia com a seleção dos dados que se pretende minerar. Após isso, os dados selecionados passam por uma etapa nomeada de pré-processamento, durante a qual diversas técnicas podem ser aplicadas visando aumentar a sua qualidade. Em seguida, os dados são transformados em um formato adequado para aplicação do algoritmo de mineração. Após essa transformação, os dados podem ser minerados e padrões podem ser extraídos. De posse desses padrões, pode-se revelar finalmente informações de alto nível (novo conhecimento), possivelmente útil e inovador.

De maneira sucinta, pode-se afirmar que a mineração dos dados é a principal etapa do processo de KDD, sendo responsável por realizar a exploração e a análise dos dados. Objetiva, portanto, encontrar padrões e gerar modelos de onde poderão ser extraídas informações úteis e previamente desconhecidas (Witten e Frank 2005). Por outro lado, não se pode desprezar o papel da etapa de pré-processamento dos dados, uma vez que esta busca maximizar a qualidade dos dados que serão utilizados pelos algoritmos de mineração. Segundo (Zhang, Zhang e Yang 2004), o pré-processamento dos dados é, apesar de menos fascinante, a tarefa mais crítica do KDD. A não-existência

[2] Acrônimo do inglês *Knowledge Discovery in Databases*.

do pré-processamento ou a sua realização mal feita pode ser responsável pelo fracasso de todo o processo de descoberta (Pyle 1999).

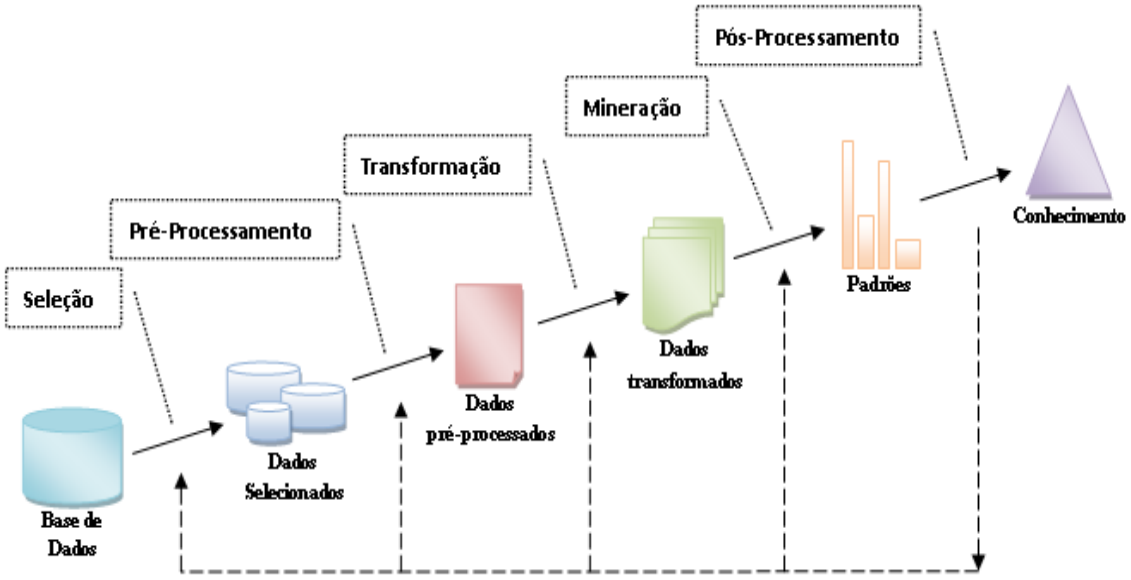

Figura 1: Esquema geral KDD (Fayyad, Gregory e Padhraic 1996)

Segundo (Zhang, Zhang e Yang 2004), pode-se destacar o valor da etapa de pré-processamento dos dados pelos seguintes aspectos:

- Os dados reais não são puros. De fato, os dados podem ser incompletos, com ruídos, ou inconsistentes, gerando desvios/vieses irreais nos padrões.
- Sistemas de mineração de alto desempenho necessitam de dados com qualidade. O tratamento dos dados, conseqüentemente, é um passo crítico para a geração de bases de dados "limpas" e menores que as originais, podendo melhorar significativamente a eficiência/eficácia dos algoritmos de mineração.
- Dados de qualidade tendem a dar origem a padrões mais úteis.

No restante deste capítulo, será dada ênfase à etapa de pré-processamento, haja vista o objetivo deste trabalho. Esta etapa envolve diversas tarefas, indo desde a remoção ou construção de atributos, adição de novas informações, remoção de instâncias e ruído ou imputação de valores faltantes (Pyle 1999) (Zhang, Yang e Zhang 2002). Existem diversas técnicas para tratar cada uma dessas tarefas, algumas delas serão apresentadas nas próximas seções.

2.1. Limpeza dos dados

A limpeza dos dados busca preencher os valores faltantes, eliminar os ruídos e corrigir inconsistências presentes nos dados (Han e Kamber 2001). Nesta seção, serão apresentados alguns dos métodos mais comuns de limpeza de dados.

2.1.1. Valores faltantes

Uma das tarefas mais utilizadas para fins de limpeza dos dados, e cerne deste trabalho, é a de tratamento de valores faltantes. Valores faltantes podem ocorrer por diversos motivos: erros de digitação; falta do conhecimento necessário para o preenchimento dos valores; os atributos podem ser opcionais; ou devido a falhas nos equipamentos de coleta. Em (Brown e Kros 2003) é apresentada a seguinte categorização para as causas da ocorrência de valores faltantes:

- Fatores operacionais: erros na entrada de dados, estimativas erradas, remoção acidental de campos de tabelas, entre outras;

- Recusa na resposta em pesquisas: alguns entrevistados podem deixar uma ou mais questões sem preenchimento. Podem se sentir constrangidos em responder a perguntas como idade, renda ou orientação sexual. Ou podem não conhecer o assunto, como por exemplo, no caso de estudantes que são argüidos acerca da carreira que desejam seguir; e

- Respostas não aplicáveis: por vezes, questões apresentadas em questionários não se aplicam aos entrevistados. Por exemplo, perguntas envolvendo ganhos de atividade rural para moradores de metrópoles, ou questões direcionadas a fumantes aplicadas a não-fumantes.

Três principais problemas estão geralmente associados à existência de valores faltantes (Barnard e Meng 1999): perda de eficiência; complicações no manuseio e análise dos dados; e análises tendenciosas ocasionadas pelas diferenças entre os dados faltantes e os dados presentes. Pode-se classificar os mecanismos que dão origem aos dados faltantes em três naturezas estatísticas (Little e Rubin 1987):

- Totalmente aleatórios (MCAR – *Missing completely at random*): diz-se que os valores faltantes são totalmente aleatórios quando a probabilidade de uma observação (amostra) ter um atributo faltante não depende direta ou indiretamente dos outros valores encontrados na base de dados, quer sejam faltantes ou não;

- Aleatórios (MAR – *Missing at random*): diz-se que os valores faltantes são aleatórios quando a probabilidade de uma observação ter um atributo faltante está de alguma forma ligada aos valores não-faltantes de algum subconjunto dos atributos e ao mesmo tempo independe dos demais valores faltantes. Ou seja, o valor faltante de um atributo pode ser estimado com base em valores preenchidos de outros atributos; e

- Não-Aleatórios (MNAR – *Missing not at random*): diz-se que os valores faltantes são não-aleatórios, quando não se encaixam como MCAR ou MAR. Ou seja, um valor faltante passa a depender de outro não-observado. Por exemplo, em uma pesquisa de salários, os entrevistados de maior renda são os que mais se recusam a responder. Isso se torna um problema quando não existe nenhuma outra variável indicadora de que o entrevistado possua uma renda alta e assim não se pode estimar qual seria o real valor do salário para esses indivíduos.

Em (Schafer e Graham 2002) tem-se ainda a indicação de uma quarta natureza: Valores fora da faixa esperada para um atributo. Os autores afirmam que os dados que se enquadram nessa situação devem ser tratados como valores faltantes. Porém, isso irá ocorrer apenas se não houver nenhuma outra tarefa de pré-processamento que trate esses valores fora da faixa esperada. MCAR é considerado o mecanismo estatístico mais comum de geração de valores faltantes (Farhangfar, Kurgan e Dy 2008); assim, foi o mecanismo assumido para este trabalho, sendo aplicado para a geração das bases de dados com dados faltantes utilizadas nos experimentos computacionais.

Existem diversas técnicas capazes de tratar os valores faltantes. A forma escolhida para tratar os dados incompletos pode ser determinante no seu sucesso. Essa escolha pode diferenciar um estudo tendencioso ou não (Twala, Cartwright e Sherpperd 2005). O mais importante aspecto da substituição de valores ausentes está na não-

distorção das características originais da amostra (Hruschka, Hruschka Jr. e Ebecken 2003).

Uma das maneiras mais simples de lidar com esse problema é remover as tuplas que contenham valores faltantes (Roth 1994). Por conta da sua simplicidade, é uma técnica bastante utilizada. Geralmente, é empregada apenas para bases em que o número de valores faltantes seja diminuto. Essa estratégia pode ser aplicada de três formas:

- Remoção completa: removem-se todas as instâncias que possuem qualquer valor faltante. O ponto negativo é que pode acarretar a perda de muitas informações;

- Remoção de colunas: são removidas da base de dados as colunas de atributos que possuam algum valor ausente. É uma abordagem não muito recomendada, uma vez que a perda de dados geralmente também é considerável; além disso, pode alterar as relações existentes entre os dados, deixando a análise tendenciosa; e

- Remoção em pares: uma variante da remoção completa. Nesta variação, uma tupla com valores faltantes pode ser utilizada desde que o atributo em questão não esteja com seu valor ausente. Caso a análise esteja sendo realizada em um atributo o qual esteja ausente na amostra, ela será ignorada. Possui a vantagem de se utilizar toda a base; porém sua implementação torna-se mais complexa.

Outra maneira simples de se trabalhar com os valores faltantes, mas sem perda de dados, seria apenas ignorar os valores faltantes (Grzymala-Busse e Hu 2001). Os valores faltantes são deixados vazios, sem nenhum tratamento. Os algoritmos de mineração, quando possível, podem utilizar-se de estratégias próprias para tratar esses valores. Um exemplo de aplicação dessa abordagem seria o algoritmo de classificação C4.5 (Quinlan 1992). O próprio algoritmo trata os valores faltantes que chegam até ele, não os considerando para alguns cálculos internos.

Em geral, o processo que vem sendo mais indicado para o tratamento de valores ausentes é o baseado em imputação, mediante o qual os valores faltantes são estimados (Little e Rubin 1987) (Farhangfar, Kurgan e Dy 2008). Na próxima subseção, serão apresentadas algumas técnicas utilizadas para este fim.

2.1.1.1. Imputação de valores faltantes

A imputação de valores faltantes alude ao processo de preenchimento dos dados ausentes de forma a gerar uma base completa. É uma alternativa bastante atrativa, pois pode gerar uma base sem perdas. Porém, uma imputação errada pode gerar mais problemas do que os resolver (Little e Rubin 1987). Na literatura atual, diversas técnicas de imputação podem ser encontradas. Abaixo são apresentadas algumas dessas técnicas:

- Substituir todos os atributos faltantes por uma constante (Grzymala-Busse e Hu 2001): baseia-se em substituir qualquer atributo faltante por um valor arbitrário. Contudo, pode levar o algoritmo a utilizar esse valor como outro qualquer, gerando distorções consideráveis;

- Substituir os atributos faltantes pela média (Grzymala-Busse e Goodwin 2005): essa técnica é utilizada apenas em atributos contínuos, sendo que o valor é calculado com base na média amostral dos valores preenchidos para o atributo em questão nas outras amostras da base. É uma técnica que pode alcançar resultados consideráveis, além de incorrer em bom desempenho computacional. Entretanto, não leva em consideração o relacionamento entre os atributos, algo muito explorado pelos algoritmos de mineração. Isso pode enviesar os dados, uma vez que estes representam apenas a tendência dos valores preenchidos;

- Substituir os atributos faltantes pelo valor mais comum (Grzymala-Busse e Goodwin 2005): visa substituir os valores faltantes pelo valor que mais ocorre, no mesmo atributo, em todas as tuplas. Pode ser aplicado tanto para atributos contínuos como para discretos. Sofre o mesmo problema da substituição pela média: ignora o relacionamento entre os atributos e pode enviesar os dados;

- Substituir por todos valores possíveis (Grzymala-Busse 1991): essa técnica se caracteriza por substituir cada valor ausente por todos os possíveis valores que o atributo pode assumir. Caso uma tupla possua mais de um valor faltante, fazem-se as alterações uma a uma, seqüencialmente, até que todos os valores faltantes estejam preenchidos. Seu ponto fraco é a grande demanda computacional;

- Imputação utilizando KNN (K-*Nearest Neighbor*) (Batista e Monard 2003) (Hruschka, Hruschka Jr. e Ebecken 2005): essa técnica busca a instância com

maior similaridade à instância que possui os valores faltantes, seguindo o preceito do paradigma de aprendizado de máquina baseado em instâncias (Aha e Kibler, 1991). Ou seja, utilizam-se os valores da instância mais próxima em consonância com uma métrica de distância para imputar os valores faltantes. Geralmente é utilizada a distância euclidiana como métrica. Uma variação é usar a média/moda dos valores das *K* instâncias mais próximas para fins de imputação, no lugar de apenas o valor de uma única instância;

- Imputação utilizando *Weighted*-KNN (Li, Deogun e Spaulding 2004): semelhante à imputação com KNN, essa técnica, por sua vez, inova ao introduzir o conceito de pesos. Leva em consideração as distâncias para com os vizinhos como pesos para se calcular o novo valor a ser imputado; os vizinhos mais próximos têm, portanto, maior relevância nesse cálculo;

- Imputação utilizando regras de associação (Bashir et al. 2009): essa técnica emprega um algoritmo de regras de associação. As regras geradas são utilizadas para preencher os valores faltantes. Caso as regras geradas não apontem para um valor a ser imputado, é empregada a técnica de imputação com KNN; e

- Imputação utilizando *K-means* (Li, Deogun e Spaulding 2004): nesta técnica usam-se agrupamentos criados por uma variante *fuzzy* do algoritmo *K-means* para imputação dos valores faltantes. Após a geração dos agrupamentos, faz-se uso do algoritmo KNN para imputação. Contudo, o KNN se limita a buscar as instâncias mais próximas no raio delimitado pelo agrupamento.

2.1.2. Ruídos

Ruídos são erros ou variações aleatórias nos valores de um atributo qualquer, levando à geração de inconsistências. Segundo (Libralon, Carvalho e Lorena 2009), a presença de ruídos em uma base pode atingir o desempenho preditivo de algoritmos de aprendizado de máquina, uma vez que pode acarretar um aumento considerável da complexidade dos modelos resultantes e do tempo necessário para sua indução.

Ruídos podem ser de diferentes tipos, tais como valores muito acima ou abaixo da média, dados rotulados de maneira incorreta, dados redundantes, entre outros. Existem diversas técnicas capazes de encontrar esses ruídos, duas das quais são relatadas abaixo:

- Agrupamentos: Cada grupo é formado por instâncias semelhantes e aquelas que não se encaixam em nenhum grupo são consideradas ruídos (*outliers*). A Figura 2 ilustra essa situação: dois agrupamentos foram formados, sendo que os pontos vermelhos seriam os ruídos; e

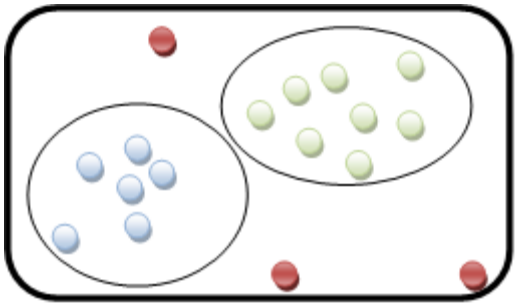

Figura 2: Agrupamentos com ruídos

- Média e desvio-padrão: Utilizando a média dos valores e o desvio-padrão, pode-se encontrar os ruídos dentro da base. Por exemplo, considerando os seguintes valores para um atributo "Idade":

  Idade = {35,50,30,200,40,33,23,1,21,39}

  Média: 47,2 / Desvio-padrão: 53,30

  Threshold: Média ± 2*Desvio Padrão (153,8 e -59,4).

  Os valores fora do *Threshold* são considerados ruídos. No exemplo apresentado, o valor '200' seria, portanto, assim considerado.

Uma das maneiras mais simples de tratar ruídos é aplicando discretização (ver próxima seção), tentando encaixar todos os valores existentes em categorias. Até mesmo os ruídos podem se inserir em alguma categoria qualquer. Por exemplo, para um atributo "Idade", criam-se categorias como "Criança, Adulto e Idoso". Desse modo, mesmo valores bem elevados podem ser classificados como da categoria "Idoso".

2.2. Transformação dos dados

A transformação dos dados envolve modificar e consolidar os dados de forma que estes fiquem mais adequados à tarefa de mineração, gerando assim melhores resultados (Han e Kamber 2001). Entre as técnicas de transformação dos dados mais comuns estão:

- Normalização: os atributos são escalados para caírem dentro uma escala especificada, geralmente pequena, como por exemplo, de 0 a 1,0;
- Agregação: trata-se de uma condensação, buscando dados mais resumidos. Por exemplo, os dados diários de vendas podem ser agregados para computar quantidades totais mensais e anuais;
- Generalização: os atributos com dados de mais baixo nível são substituídos por informações de mais alto nível. Por exemplo, se uma base possui o atributo "Rua", esse atributo poderia ser generalizado para "Cidade" ou "País";
- Discretização: também considerada como uma espécie de generalização. Consiste na troca dos atributos contínuos por valores discretos. Por exemplo, um atributo contínuo "Idade" pode ser substituído por valores discretos tais como: "criança", "jovem", "adulto" e "idoso"; e
- Construção de atributos: novos atributos são construídos tomando como base os já existentes, geralmente mediante combinações (lineares ou não) entre eles.

2.3. Redução dos dados

As técnicas de redução dos dados buscam simplificar as bases de dados na esperança de que os algoritmos de mineração possam chegar a resultados mais relevantes e com menor tempo de processamento. Apesar dos ganhos evidentes, deve-se observar que a redução dos atributos pode causar grandes distorções em relação aos dados reais, o que pode gerar resultados irreais. Dentre as principais técnicas para

redução dos dados estão a seleção de atributos e a seleção de instâncias, as quais são brevemente discutidas nas próximas subseções.

2.3.1. Seleção de atributos

A seleção de atributos visa refinar a lista de atributos de uma base para remover eventuais ruídos e atributos irrelevantes, bem como eliminar possíveis ambigüidades. É selecionado um subconjunto dos atributos de uma base de dados de modo a escolher os atributos mais relevantes. A utilização de atributos irrelevantes pode ser prejudicial ao processo de mineração, podendo diminuir a exatidão dos padrões encontrados ou mesmo tornar os resultados mais confusos e menos expressivos.

Quanto maior o número de atributos aumenta-se também o número de instâncias necessárias para se encontrar bons padrões dentro das bases de dados (Yang e Honavar 1997); assim sendo, a seleção de atributos torna-se útil não só quando as bases possuem um grande número de atributos, como também quando as bases possuem um pequeno número de instâncias. Diminuindo o tamanho da base de dados, consegue-se ainda melhorar o desempenho computacional dos algoritmos de mineração. Pode-se observar, de maneira simplificada, três tipos de algoritmos para realizar a seleção de atributos (Boz 2002): algoritmos exponenciais, seqüenciais e aleatórios.

Algoritmos exponenciais envolvem buscas exaustivas, em que são feitas todas as combinações possíveis com os valores dos atributos disponíveis. Pode se tornar computacionalmente inviável devido ao tempo de processamento, que cresce exponencialmente com o número de atributos utilizados.

Algoritmos seqüenciais visam selecionar um subconjunto dos atributos, fazendo os agrupamentos seqüencialmente, podendo ser para frente ou para trás. Na busca seqüencial para frente, primeiro seleciona-se um único atributo, o mais bem avaliado isoladamente e, após isso, este atributo é avaliado em par com todos os outros, e assim por diante. A cada nova iteração, um novo atributo é adicionado à solução. Na busca seqüencial para trás, o subconjunto inicial é composto por todos os atributos disponíveis, a cada iteração um dos atributos é removido, continuando assim até se encontrar a melhor solução. Assim como os algoritmos exponenciais, os algoritmos seqüenciais podem se tornar computacionalmente inviáveis devido ao tempo de processamento que necessitam.

Os algoritmos aleatórios abarcam todos os métodos que selecionam subconjuntos de atributos via busca aleatória; um exemplo deste tipo de busca envolve os algoritmos genéticos. Uma das vantagens dos algoritmos aleatórios está no fato de não se ter que necessariamente utilizar todos os atributos disponíveis em sua busca, podendo reduzir o tempo total de processamento.

Pode-se aplicar a seleção de atributos de acordo com duas abordagens: *filter* e *wrapper* (Kohavi e John 1997). Na abordagem *filter* (filtro), a seleção de atributos é independente do algoritmo de mineração de dados que será aplicado. Em contrapartida, a abordagem *wrapper* (de empacotamento) utiliza o resultado do algoritmo de mineração para determinar o quão bom um subconjunto de atributos é. Nessa abordagem, o método de seleção gera subconjuntos (soluções-candidatas) iterativamente e os avalia com base no desempenho do algoritmo de mineração. Esse processo se repete até que uma condição de parada seja satisfeita.

2.3.2. Seleção de instâncias

A seleção de instâncias visa obter uma representação reduzida da base de dados, removendo tuplas irrelevantes e sem causar grandes perdas das informações originais. Seleciona-se um subconjunto das tuplas da base de dados, de modo a deixá-la apenas com os dados relevantes para os algoritmos de mineração.

Assim como, quanto mais atributos são utilizados, maior é o número de instâncias necessárias para se obter bons resultados, o contrário também geralmente é verdade. Ou seja, quanto maior o número de instâncias, maior o número de atributos necessários para alcançar melhores resultados. Isso implica que a seleção de instâncias seja uma boa tarefa a ser realizada em bases que possuam poucos atributos.

O uso de um elevado número de instâncias pode degradar em muito o desempenho computacional dos algoritmos de mineração. Muitas dessas instâncias podem chegar a reduzir a exatidão dos modelos gerados, caso possuam muitos ruídos ou dados irrelevantes. Assim como para seleção de atributos, pode-se aplicar três tipos de algoritmos para realizar a seleção de instâncias: os algoritmos exponenciais, seqüenciais

e os aleatórios. Todos seguem a mesma lógica discutida anteriormente, porém para as tuplas das bases de dados.

2.4. Conclusão

Neste capítulo, foram apresentados conceitos importantes relacionados à etapa de pré-processamento de dados no KDD. Foram discutidas as tarefas de limpeza dos dados, transformação dos dados e redução dos dados. A limpeza dos dados atua na retirada das inconsistências ou ruídos existentes na base e preenchimento dos valores faltantes. A transformação dos dados visa consolidar os dados existentes de maneira que assumam um formato mais adequado para a mineração. Já a redução tem como objetivo diminuir a base de dados a fim de deixá-la apenas com dados relevantes para a mineração, sem perder sua essência inicial.

No próximo capítulo, serão abordados os algoritmos heurísticos bio-inspirados utilizados neste trabalho para realizar especificamente o tratamento de valores faltantes.

3. ALGORITMOS HEURÍSTICOS BIO-INSPIRADOS

Serão abordados neste capítulo alguns conceitos relacionados aos algoritmos heurísticos bio-inspirados utilizados neste estudo: algoritmos genéticos, otimização por enxame de partículas e co-evolução cooperativa. Os três algoritmos pertencem a um novo ramo da computação intitulado de Computação Natural. A Computação Natural envolve a criação de novas abordagens computacionais inspiradas em idéias e fenômenos da natureza. Segundo (de Castro 2007), pode-se dividir a Computação Natural em três ramos:

- Computação inspirada na natureza: inclui algoritmos ou técnicas computacionais desenvolvidas com base em modelos biológicos, tais como as redes neurais artificiais, a computação evolutiva, a inteligência de enxame, e os sistemas imunológicos artificiais;
- Síntese de formas de comportamentos naturais: envolve a utilização de mecanismos computacionais voltados para a síntese de comportamentos naturais, bem como de padrões e processos biológicos. Nesse contexto, se insere os métodos computacionais baseados na geometria fractal. Abarca também estudos a respeito da vida e de organismos artificiais; e
- Computação via mecanismos naturais: trata-se de um novo paradigma de computação segundo o qual dispositivos naturais (por exemplo, cadeias de DNA e computadores quânticos) são utilizados como novas arquiteturas ou estruturas de dados computacionais.

O três algoritmos empregados neste trabalho pertencem ao ramo da computação inspirada na natureza. Podem ser considerados como meta-heurísticas populacionais voltadas para a resolução de problemas de otimização contínua ou discreta. Segundo

(Blum e Roli 2003), meta-heurísticas são estratégias de alto nível destinadas a explorar o espaço de busca associado a um problema, podendo se utilizar para tanto de diferentes métodos heurísticos de mais baixo nível. Os autores sintetizam algumas características gerais exibidas por diferentes classes de meta-heurísticas, dentre as quais as de gênero bio-inspirado:

- São estratégias gerais que guiam o processo de busca;
- Têm como objetivo explorar o espaço de busca de forma eficiente, a fim de encontrar uma solução ótima;
- As técnicas que compõem os algoritmos meta-heurísticos vão desde simples buscas locais até complexos processos de aprendizagem;
- São aproximativos e, geralmente, não-determinísticos;
- Podem possuir mecanismos para evitar o confinamento prematuro em áreas específicas do espaço de busca; e
- São bem flexíveis e não tratam apenas de um problema específico.

Nas próximas seções, serão apresentados os conceitos relativos a cada um dos algoritmos heurísticos bio-inspirados utilizados neste estudo.

3.1. Algoritmos genéticos

Os algoritmos genéticos (GA) são métodos estocásticos de busca e otimização inspirados nos mecanismos de evolução das populações de seres vivos. Seguem os princípios da Seleção Natural e sobrevivência do mais apto, propostos por Charles Darwin: “Quanto melhor um indivíduo se adaptar ao seu meio ambiente, maior será sua chance de sobreviver e gerar descendentes”.

São geralmente implementados como uma simulação computacional, na qual uma população de possíveis soluções (indivíduos) evolui a fim de encontrar a melhor solução para um dado problema. Cada indivíduo possui a representação de um cromossomo. Sobre esses indivíduos são aplicados operadores como a seleção, a reprodução e a mutação. O pseudocódigo de um algoritmo genético típico é apresentado na Tabela 1 (Linden 2006).

Tabela 1: Pseudocódigo de um algoritmo genético típico

```
Seja S(t) a população inicial na geração t
t←0
iniciar S(t)
avaliar S(t)
enquanto (critério de parada não satisfeito) faça
    t←t+1
    selecionar P(t) (conjunto de pais) a partir de S(t-1)
    aplicar recombinação sobre P(t) dando origem a F(t)
    aplicar mutação sobre F(t)
    avaliar F(t)
    selecionar dentre P(t) e F(t) os indivíduos para S(t)
fim enquanto
```

O primeiro passo é a construção de uma população inicial de indivíduos, freqüentemente gerada aleatoriamente. Cada indivíduo recebe uma avaliação (chamada de aptidão ou *fitness*) que reflete a qualidade da solução que ele representa. Após isso, o algoritmo entra em um laço que representa as gerações do processo evolutivo. No início de cada geração, o operador de seleção para reprodução é aplicado. Geralmente, nessa seleção, os indivíduos que possuem maior grau de aptidão são mais suscetíveis a serem escolhidos como pais. Isso não quer dizer que os demais não possam ser escolhidos também, já que se trata geralmente de um operador estocástico. Esses indivíduos selecionados são recombinados (via operador de cruzamento (*crossover*)), processo que simula a reprodução da população. Com os novos indivíduos gerados, eles passam por uma etapa de mutação e são então avaliados. De posse dos filhos, são selecionados os indivíduos que farão parte da população na próxima geração (seleção para substituição), dentre pais e filhos. O laço é reiniciado até que a condição de parada seja atingida e o algoritmo termine.

O algoritmo genético depende de alguns parâmetros de controle, usados desde para a criação da primeira população como durante o processo de evolução. Os operadores genéticos e de seleção utilizados são geralmente estocásticos e sua aplicação é direcionada por meio desses parâmetros de controle, os quais são usualmente previamente definidos. Alguns dos parâmetros utilizados são: o tamanho da população, a taxa de seleção, a taxa de reprodução e a taxa de mutação. Um esquema básico de um algoritmo genético pode ser observado na Figura 3 (Linden 2006).

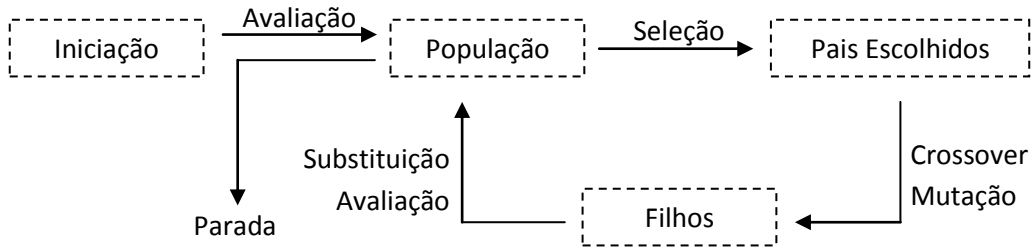

Figura 3: Esquema básico de um algoritmo genético típico

A representação dos indivíduos é um passo fundamental para os algoritmos genéticos. Essa representação irá mapear as informações de um problema que se deseja resolver em informações tratáveis pelo algoritmo. Quanto melhor a representação dos indivíduos se adequar ao problema, maiores as chances de se obter melhores resultados (Linden 2006). Em (Linden 2006), são destacadas três características importantes para uma boa representação: ela deve ser a mais simples possível; caso existam soluções indesejáveis, elas não deveriam ser representáveis; e, caso o problema imponha alguma condição, esta deve estar implícita na representação utilizada. Existem diversas maneiras de se representar os indivíduos, as mais comuns sendo:

- Representação binária: nesta, os cromossomos nada mais são do que uma seqüência de *bits*. Cada gene do cromossomo pode assumir apenas os valores zero e um. Foi introduzida por Holland (Holland 1975) e, por sua simplicidade e aplicabilidade em diversos problemas, é uma das representações mais utilizadas;
- Representação numérica: difere da representação binária pelo fato de que cada gene não é representado apenas por um *bit* (zero ou um), mas sim por um número (contínuo ou discreto) qualquer;
- Representação categórica: utilizada quando se deseja representar valores que são categóricos. Cada gene pode assumir um valor dentro de um conjunto de valores predefinidos; e
- Representação híbrida: utilizada quando o problema não pode ser resolvido utilizando apenas um tipo de representação. Utilizam-se, daí, dois ou mais tipos de representações em um mesmo cromossomo.

O operador de seleção também pode ser aplicado de diferentes formas, sempre buscando dar maior prioridade aos indivíduos mais aptos; porém, sem eliminar as chances dos menos aptos. Algumas implementações comuns são:

- Roleta: visa criar uma roleta virtual na qual cada indivíduo recebe uma porção proporcional à sua aptidão. Simula-se uma rodada na roleta e o indivíduo associado ao ponto em que a roleta pára é selecionado;
- Ordenação: o método de ordenação possui a mesma filosofia do método da roleta. A diferença está na formação da roleta. A porção de cada indivíduo não é proporcional à sua aptidão, mas sim à sua classificação. É indicado para problemas nos quais alguns indivíduos se sobressaem em muito do restante da população; e
- Torneio: neste método, um número *N* de indivíduos é escolhido aleatoriamente. Esses são comparados entre si e o de maior aptidão é selecionado. Esse processo se repete até que o número de indivíduos desejado seja alcançado.

O operador de mutação é estocástico e também existem várias estratégias para realizá-lo. Sua implementação mais comum consiste em aplicar uma pequena alteração em um gene qualquer de um indivíduo. A mutação objetiva retirar o algoritmo de máximos ou mínimos locais e permitir com que o algoritmo alcance, teoricamente, qualquer região do espaço de busca. Já o operador de recombinação (*crossover*) realiza a reprodução entre os indivíduos. Consiste na troca de materiais genéticos entre dois representantes da população. O método mais comum de cruzamento é realizado efetuando a união de parte dos genes dos cromossomos-pais. É definido um ponto de corte qualquer nos cromossomos, e a parte direita de um pai é conectada com a parte esquerda do outro para gerar um novo filho. O mesmo ocorre com as outras partes para geração de outro filho.

3.2. Otimização por enxame de partículas

Otimização por enxame de partículas (PSO[3]) é uma meta-heurística populacional

[3] Acrônimo do inglês *Particle Swarm Optimization*.

pertencente à subárea de pesquisa da Computação Natural conhecida como Inteligência Coletiva (Kennedy, Eberhart e Shi 2001). Baseia-se na simulação do comportamento coletivo exibido por algumas espécies de animais (Kennedy e Eberhart 1995), tais como aquele observado em um bando de pássaros ou em um cardume de peixes. Pode-se observar a grande sincronia com que um cardume se movimenta, sempre buscando um objetivo comum ao grupo, e sem se chocarem, muitas vezes, utilizando informações de experiências já vividas para guiar seus movimentos futuros. Fazendo analogia a um cardume, no PSO, cada peixe é representado por uma partícula. Cada partícula representa uma solução-candidata para um problema de busca/otimização. Essas partículas se movimentam pelo espaço de busca procurando por soluções ótimas. O movimento de cada partícula é baseado nas experiências vividas até aquele momento por ela e nas experiências vividas pelo restante do grupo.

Cada partícula é composta por um vetor-solução de dimensão *d*, $X_i = (x_{i1}, x_{i2}, \dots, x_{id})$, e um vetor de velocidade, $V_i = (v_{i1}, v_{i2}, \dots, v_{id})$. Possui ainda uma referência para o melhor local já visitado por ela, $P_b = (p_{b1}, p_{b2}, \dots, p_{bd})$, uma referência para o melhor local já visitado pelas partículas que compõem sua vizinhança, $P_g = (p_{g1}, p_{g2}, \dots, p_{gd})$, além do seu valor de *fitness*. A movimentação de cada partícula é guiada por sua velocidade, pelo melhor ponto já visitado por ela e pela melhor posição já visitada pela vizinhança. Existem duas versões de implementação do PSO (Kennedy e Eberhart 2001): a versão local, que utiliza o conceito de vizinhança como sendo um conjunto de indivíduos próximos (logicamente ou fisicamente); e a versão global, em que a vizinhança engloba toda a população.

Por ser um método simples, de fácil implementação e baixo custo computacional (Eberhart, Simpson e Dobbins 1996), PSO tem sido foco de vários estudos. O pseudocódigo do PSO padrão é apresentado na Tabela 2.

O vetor de velocidade de cada partícula é recalculado a cada iteração, seguindo a equação:

$$v_i(t+1) = w v_i(t) + c_1 \emptyset_1 \left(p_{bi}(t) - x_i(t)\right) + c_2 \emptyset_2 \left(p_g(t) - x_i(t)\right), \qquad (1)$$

na qual w representa o peso de inércia, c_1 e c_2 são os parâmetros cognitivo e social, respectivamente, e $\emptyset_1$ e $\emptyset_2$ são variáveis aleatórias no intervalo [0,1]. A função do peso

de inércia (w) é controlar o impacto do valor da velocidade em gerações anteriores sobre o seu valor na geração atual. Valores altos facilitarão a exploração de uma maior área do espaço de busca, mas podem comprometer o tempo para convergência e a utilização de recursos computacionais; por outro lado, valores pequenos facilitarão o refinamento da solução em regiões de ótimos locais, mas podem implicar em uma convergência prematura (Shi e Eberhart 1998a) (Shi e Eberhart 1998b). Os valores de c_1 e c_2 contribuem para uma rápida convergência, para a fuga de ótimos locais e para o refinamento do resultado final.

Tabela 2: Pseudocódigo do PSO padrão

```
t ← 0
Seja P(0) a população na geração inicial
Avaliar as partículas em P(0)
P_g ← 0
para cada partícula i faça
      Pb_i ← 0
fim para
enquanto (critério de parada não satisfeito) faça
      t ← t+1
      para cada partícula i faça
            atualizar velocidade de P_i(t)
            atualizar posição de P_i(t)
            se Pb_i(t) > Pb_i(t-1) então atualizar Pb_i(t)
      fim para
      se P_g(t) > P_g(t-1) então atualizar P_g(t)
fim enquanto
```

Tendo a velocidade calculada, a partícula é movimentada para sua próxima posição seguindo a equação:

$$x_i(t+1) = x_i(t) + v_i(t). \quad (2)$$

Na equação do cálculo da velocidade, não existe nenhum mecanismo que limite seu valor; por isso, muitas vezes é utilizado um parâmetro V_{max} para este fim, evitando que a partícula seja lançada para fora do espaço de busca. Este valor deve ser bem calibrado, pois, caso fique muito alto, ele irá possibilitar que as partículas saiam do

espaço de busca, ao passo que valores pequenos poderão resultar em exploração insuficiente do espaço de busca.

Em uma variação proposta por (Clerc e Kennedy 2002), um novo parâmetro (χ), conhecido como coeficiente de constrição, foi adicionado à Equação (1). Esse parâmetro tem a função de fazer com que as partículas recebam valores menores para os seus vetores de velocidade a cada geração, evitando também que elas escapem do espaço de busca. A nova equação pode ser observada em (3):

$$v_i(t+1) = \chi\left(v_i(t) + c_1\emptyset_1\left(p_{bi}(t) - x_i(t)\right) + c_2\emptyset_2\left(p_g(t) - x_i(t)\right)\right). \quad (3)$$

É importante ressaltar que, na Equação (3), quando o coeficiente de constrição χ é utilizado, não se deve usar o peso de inércia, e vice-versa. Os outros parâmetros são os mesmos utilizados na Equação (1). Na Equação (4), temos o cálculo usualmente adotado para o fator de constrição, sendo que $\varphi = c_1 + c_2, \varphi > 4$ (Clerc e Kennedy 2002):

$$\chi = \frac{2}{2-\varphi-\sqrt{\varphi^2-4\varphi}}. \quad (4)$$

3.3. Co-evolução cooperativa

A utilização de algoritmos genéticos pode ser muito custosa em termos computacionais, principalmente nos casos em que o problema possuir um grande número de variáveis ou quando é necessário tratar mais de um problema simultaneamente. Uma estratégia indicada para esses casos é a decomposição desses problemas em problemas menores. É uma estratégia bastante utilizada para resolução de problemas complexos na área da computação, nomeada de "dividir-para-conquistar". Baseia-se em resolver as partes menores do problema separadamente, e, posteriormente, agrupar as soluções para gerar a solução final (Coelho 2004).

Neste contexto, a abordagem de co-evolução cooperativa, como proposta inicialmente por (Potter e De Jong 1994), tenta aplicar a estratégia de "dividir-para-conquistar" para resolução de problemas de otimização utilizando algoritmos genéticos. Baseia-se na criação de um ambiente (ecossistema) no qual dois ou mais algoritmos

genéticos (espécies) possam evoluir cooperativamente, colaborando entre eles para alcançar um objetivo comum. Nesse sentido, essa abordagem combina aspectos evolutivos e ecológicos em um só arcabouço conceitual e vem sendo aplicada em diversos trabalhos, tais como os de (Puppala, Sen e Gordin 1998), (Wiegand, Liles e De Jong 2001), (Roberts e Claridge 2004), (Potter e De Jong 1999;2000) e (Coelho 2004).

A arquitetura utilizada neste estudo foi baseada naquela exposta no trabalho de (Potter e De Jong 2000), envolvendo também a extensão de memória compartilhada proposta por (Puppala, Sen e Gordin 1998). Esta arquitetura pode ser observada na Figura 4. A arquitetura contempla a existência de *N* espécies compartilhando um mesmo ambiente e evoluindo cooperativamente.

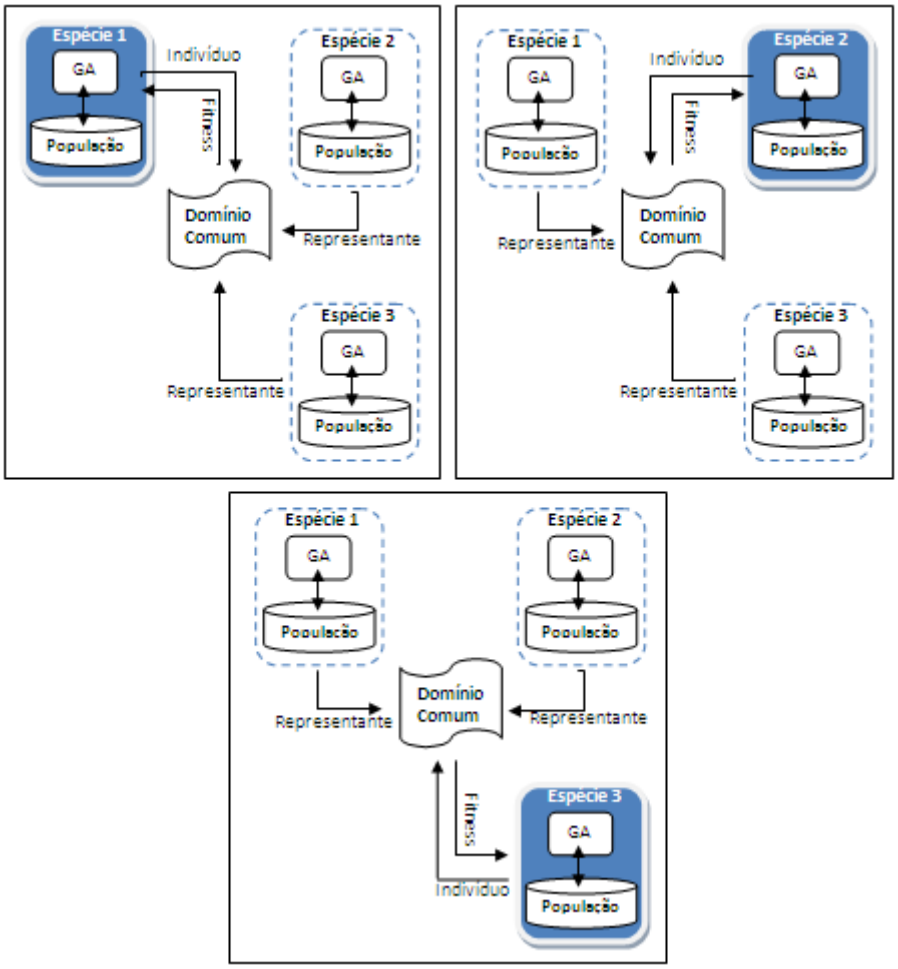

Figura 4: Arquitetura de co-evolução

No centro, existe um domínio comum, no qual as espécies interagem para fins de cálculo do valor de *fitness*. A avaliação de um indivíduo necessita da participação de pelo menos um indivíduo de cada uma das outras espécies. O modo como esses indivíduos das outras espécies são obtidos pode ocorrer de diferentes maneiras, sendo que três delas são descritas a seguir (Wiegand, Liles e De Jong 2001):

- Aleatória: realiza a combinação escolhendo aleatoriamente qualquer indivíduo de cada uma das outras espécies;
- Melhor indivíduo: faz a combinação utilizando os melhores indivíduos de cada uma das outras espécies; e
- Utilizando memória (*buffer*): no domínio, é guardado um número *X* das melhores combinações já realizadas. O indivíduo que está sendo avaliado substitui o indivíduo da sua espécie correspondente em todas as combinações existentes na memória. O melhor resultado será atribuído como *fitness* do indivíduo em avaliação. Caso o resultado obtido pela nova combinação seja melhor do que o resultado de alguma das combinações existentes na memória, a nova combinação irá entrar na memória e a pior combinação será retirada (Puppala, Sen e Gordin 1998).

Todas essas estratégias possuem pontos positivos e negativos. A estratégia aleatória permite explorar bastante o espaço de busca, podendo evitar cair em máximos ou mínimos locais; contudo, possui menos chances de alcançar o melhor resultado. A estratégia de escolher os melhores indivíduos tem grandes chances de alcançar um bom resultado; entretanto, pode levar mais facilmente ao confinamento em máximos ou mínimos locais. Segundo (Puppala, Sen e Gordin 1998), a abordagem utilizando memória compartilhada é mais interessante em termos de eficiência e eficácia, pois permite realizar um maior número de avaliações em regiões promissoras, ao mesmo tempo em que fugindo de máximos e mínimos locais. No entanto, é a estratégia de pior desempenho computacional.

Quanto às espécies que participam do processo, pode-se classificar a co-evolução em dois tipos (Coelho 2004): homogênea e heterogênea. Na co-evolução homogênea trabalha-se apenas com um tipo de espécie, ou seja, tratando apenas um tipo de problema. Nesta, o problema é dividido em duas ou mais partes e essas partes são

distribuídas entre as espécies. Na co-evolução heterogênea trabalha-se com dois ou mais tipos de espécies diferentes. Cada espécie trata de um problema distinto. Neste caso, o objetivo é a melhor solução para todos os problemas em conjunto.

3.4. Conclusão

Foram apresentados, neste capítulo, conceitos relativos a três algoritmos heurísticos bio-inspirados que serviram de investigação neste trabalho: algoritmos genéticos, PSO e co-evolução cooperativa. Enquanto os algoritmos genéticos utilizam-se de operadores baseados na natureza, seguindo as teorias de Darwin, o PSO é baseado na simulação do comportamento coletivo exibido por algumas espécies de animais, como peixes, pássaros, lobos, etc. Analisou-se ainda neste capítulo a abordagem de co-evolução cooperativa. Nesse contexto, apresentou-se a arquitetura que permite a utilização de um número arbitrário de espécies (associadas a diferentes instâncias de algoritmos genéticos) para resolução de um ou mais problemas, de forma concorrente.

No próximo capítulo, serão apresentados e discutidos os experimentos conduzidos com esses algoritmos bio-inspirados para o tratamento automático de atributos faltantes. Serão analisados os resultados obtidos pela abordagem de imputação bio-inspirada, comparando-os com aqueles obtidos via técnicas tradicionais de imputação.

4. EXPERIMENTOS COMPUTACIONAIS

Neste capítulo, serão apresentados e discutidos os experimentos realizados neste trabalho para fins de tratamento de dados faltantes. Ademais, serão delineados aspectos e componentes relativos às técnicas de imputação bio-inspiradas adotadas neste estudo.

Os testes experimentais foram realizados com o intuito principal de mensurar, em termos de eficiência e eficácia, a abordagem de imputação por otimização utilizando heurísticas bio-inspiradas. O objetivo central é o de mensurar o quanto essa abordagem é capaz de melhorar o desempenho de diferentes tipos de classificadores. A eficácia de uma técnica de imputação é medida pela taxa média alcançada no processo de validação cruzada (10-*fold cross-validation*) (Witten e Frank 2005), como descrito à frente, ao passo que a sua eficiência é dada pelo tempo computacional total gasto durante o pré-processamento.

Os experimentos foram conduzidos utilizando bases de dados do repositório UCI (Asuncion e Newman 2007). As bases foram escolhidas de forma a se obter diferentes números de amostras, de distribuições das classes e de números/tipos de atributos. As bases utilizadas e suas propriedades estão descritas na Tabela 3. A tabela se encontra organizada da menos para a mais complexa base.

Tabela 3: Bases de dados utilizadas nos experimentos
NC: número de classes; NT: número de tuplas; NA: número de atributos; TA: tipo do atributo (Numérico e Discreto)

Nome	*NC*	*NT*	*Distribuição das Classes*	*NA*	*TA*
IRIS	3	150	(33,3%;33,3%;33,3%)	4	N
ZOO	7	101	(40,6;19,8%;4,9%;12,9%;4,0%;7,9%;9,9%)	16	D&N
DIABETES	2	768	(65,1%;34,9%)	8	N
BCD699	2	699	(65,6%;34,5%)	9	N
CAR	4	1728	(70,0%;22,2%;4,0%;3,8%)	6	D
DERMATOLOGY	6	366	(30,6%;16,7%;19,7%;13,4%;14,2%;5,5%)	33	D&N

Originalmente, com exceção da *BCD699* e *Dermatology*, essas bases não possuíam valores faltantes. As bases foram alteradas, de maneira controlada, para inseri-los. Esta estratégia permite ter um maior controle sobre o impacto dos dados faltantes, tornando os experimentos mais justos e reprodutíveis. Esta mesma estratégia também foi adotada nos trabalhos apresentados em (Batista e Monard 2003) e (Liu e Lei 2005), nos quais os dados faltantes são inseridos de maneira completamente aleatória.

Foram utilizadas quatro taxas para inserção de valores faltantes: 5%, 10%, 20% e 33%. Cada base original foi inicialmente particionada em duas: validação cruzada (80%) e teste (20%). Manteve-se a distribuição das classes em cada partição. Esse processo de particionamento foi realizado três vezes para cada base, gerando assim três pares para cada conjunto de dados. Posteriormente, para cada um dos pares, foram enxertados os valores faltantes, seguindo as quatro taxas supracitadas. Esse passo foi repetido cinco vezes para cada taxa, conforme representado na Figura 5. No total, foram geradas 60 diferentes bases derivadas para cada uma das seis bases originais.

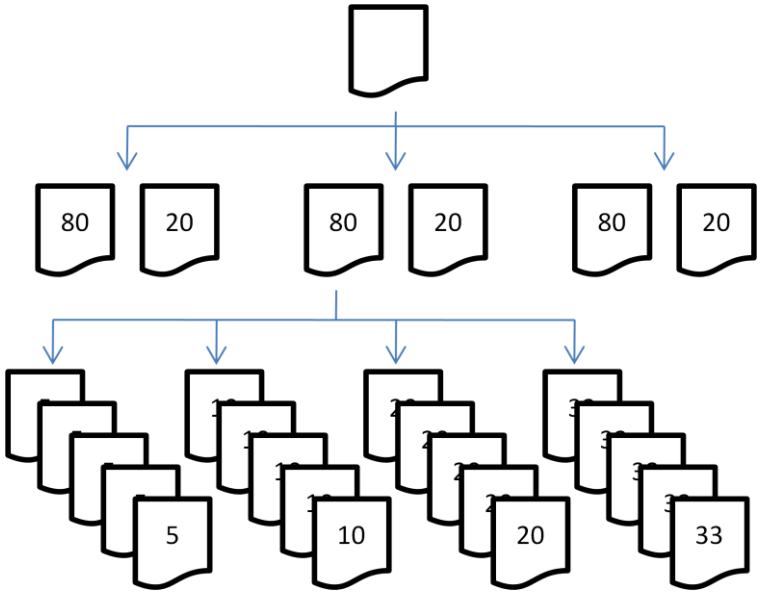

Figura 5: Geração das bases para o experimento

Nos testes, foram empregados cinco algoritmos populares de classificação disponíveis na ferramenta *Weka* (Witten e Frank 2005). Cada algoritmo representa um diferente tipo de classificador/indutor: um algoritmo baseado em árvores de decisão, o

J48; um algoritmo baseado na estatística bayesiana, o *Naïve Bayes*; um outro baseado em regras de decisão, o PART; redes neurais *feedforward* com neurônios escondidos usando função de ativação do tipo base radial (RBF); e, por fim, um algoritmo concordante com o paradigma de aprendizado baseado em instâncias, o IBK.

O algoritmo J48 é uma implementação em Java do algoritmo de árvore de decisão C4.5 (Quinlan 1992). Trata-se de uma versão melhorada do algoritmo C4.5, a versão anterior a C5.0, que é comercial. Gera um modelo em árvore via cálculo da entropia dos atributos, sendo que cada folha da árvore resultante representa uma classe e os demais nós especificam testes condicionais a serem realizados em um determinado atributo. Cada possível valor de um atributo gera um ramo na árvore, conforme o exemplo apresentado na Figura 6.

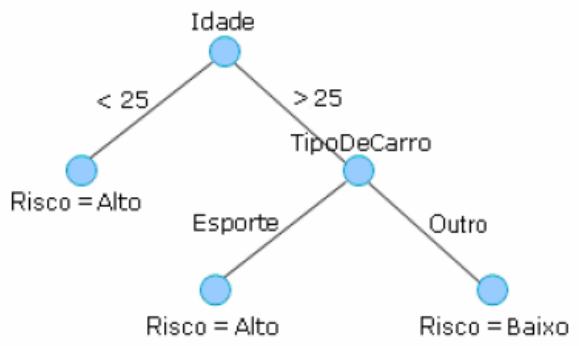

Figura 6: Exemplo de árvore de decisão

Redes neurais RBF constituem um tipo popular de rede neural de três camadas (Harpham, Dawson e Brown 2004). Cada nó da camada intermediária representa um agrupamento de pontos no espaço de entrada. Sua saída para uma instância apresentada na camada de entrada depende apenas da distância entre o seu centróide do agrupamento representado e a própria instância de entrada. Quanto maior a distância, maior a força de ativação do neurônio (Haykin 1998). A camada intermediária utiliza uma função de ativação de base radial para converter a distância em uma medida de similaridade. A camada de saída, por sua vez, é responsável por combinar linearmente as saídas geradas na camada intermediária e gerar a estimação final. A Figura 7 ilustra uma configuração típica dessa rede neural.

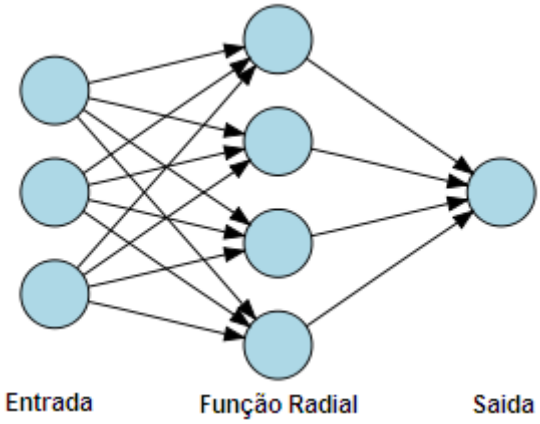

Figura 7: Exemplo de rede neural RBF

O algoritmo *Naïve Bayes* (George e Langley 1995) é de cunho probabilístico, sendo baseado na estatística bayesiana. O algoritmo assume que os atributos sejam condicionalmente independentes; ou seja, a informação representada por um atributo não é informativa sobre a de nenhum outro. Diz-se que X é condicionalmente independente de Y, dado Z, se a distribuição de probabilidades de X é independente do valor de Y, dado o valor de Z, isto é:

$$(\forall x_i, y_j, z_k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k). \quad (5)$$

O algoritmo IBK (Aha e Kibler 1991) gera modelos estatísticos baseado no *K-nearest neighbors* (KNN), sendo classificado como um algoritmo de aprendizado "preguiçoso", haja vista que não gera nenhum modelo a partir dos dados (utiliza apenas as instâncias em si). Este algoritmo assume que instâncias similares possuam a mesma classificação e depende fortemente da métrica usada como distância.

Por fim, o algoritmo PART (Frank e Ian 1998) é um indutor de modelos baseados em regras de decisão. Utiliza internamente o C4.5, constrói árvores de decisão parciais a cada iteração e transforma a melhor folha desta árvore em uma regra. Após escolher a melhor folha, o algoritmo retira todas as tuplas que se encaixem na regra gerada pela folha para gerar uma nova árvore, e, por conseguinte, uma nova regra. Esse processo é realizado iterativamente até que não se tenha mais tuplas a serem utilizadas. Como não usa a base inteira para gerar as árvores, este algoritmo geralmente incorre em um bom desempenho computacional quando da indução do modelo de classificação.

4.1. Técnicas de imputação adotadas

Neste trabalho, foram empregadas cinco diferentes técnicas para tratamento de valores faltantes. Duas tradicionais: imputação pela média/moda e imputação utilizando o algoritmo KNN. E três bio-inspiradas: imputação utilizando GA, imputação utilizando PSO e imputação utilizando co-evolução cooperativa.

As técnicas de imputação tradicionais utilizadas seguem uma abordagem similar aos métodos de filtro (*filter*) empregados para seleção de atributos (vide Subseção 2.3.1). Essa abordagem é independente do modelo de classificação que será gerado posteriormente. Essa independência para com o modelo faz com que a abordagem não garanta bons resultados para um classificador qualquer. Porém, tem a vantagem de incorrer em bom desempenho computacional, além do fato de que, uma vez realizada, a base resultante pode ser usada para a indução de diferentes classificadores em paralelo. Para a imputação utilizando o algoritmo KNN, foram aplicados três valores de *K* (número de vizinhos): um, três e cinco. Ambas as técnicas dessa classe são não-estocásticas; por isso, a sua eficácia é mensurada com base em apenas uma execução.

Para a abordagem de imputação via otimização numérica, foram empregadas três técnicas: imputação utilizando GA, imputação utilizando PSO e imputação utilizando co-evolução cooperativa. As três técnicas se assemelham aos métodos do tipo *wrapper*, também empregados para fins de seleção de atributos (vide Subseção 2.3.1). Trata-se, portanto, de uma abordagem de imputação diretamente ligada ao modelo. Para cada conjunto de valores imputados a ser avaliado, é treinado um classificador, representando o modelo, e seu desempenho serve como medida de qualidade. Essa abordagem não garante que um bom resultado de imputação para um dado modelo possa ser reaproveitado para outro, inviabilizando a indução de vários modelos em paralelo.

Para cada técnica de imputação bio-inspirada foram realizadas 10 execuções. Esse número foi necessário devido à característica estocástica dessas técnicas. A média das taxas obtida através da validação cruzada das dez simulações representa a eficácia da abordagem; já a média do tempo de processamento representa sua eficiência.

Em um trabalho recente (Farhangfar, Kurgan e Dy 2008), é apresentado um estudo comparativo sobre imputação de valores faltantes. Nele, é apresentado um resumo teórico de outros quatro trabalhos que se inserem no mesmo contexto, sendo

eles: (Acuna e Rodriguez 2004), (Batista e Monard 2003), (Grzymala-Busse e Hu 2001) e (Mundfrom e Whitcomb 1998). A Tabela 4 traz uma visão resumida desses trabalhos como apresentados em (Farhangfar, Kurgan e Dy 2008); nesta tabela, também adicionamos as configurações experimentais seguidas pelo presente trabalho. Pode-se assim perceber como o estudo comparativo conduzido aqui se posiciona, em termos de abrangência e significância, em relação ao estado-da-arte em tratamento de valores faltantes.

Nas próximas subseções, serão descritos detalhes de concepção e implementação relativos às técnicas de imputação bio-inspiradas.

Tabela 4: Tabela comparativa de trabalhos experimentais relacionados

Referência		Acuna e Rodriguez 2004	Batista e Monard 2003	Grzymala-Busse e Hu 2000	Mundfrom e Whitcomb 1998	Farhangfar, Kurgan e Dy 2008	Este trabalho
Métodos	Número Nome do Método	3 Média, Mediana, KNN	3 Média, Moda, KNN	5 Moda, Baseado no C4.5, Todos os valores, Baseado no LEM2, Event Covering	3 Média, Regressão, Hot Deck	6 Média, Hot Deck, Framework com Hot Deck, Naive Bayes, Framework com Naive Bayes, Regressão polinomial múltipla	7 Média, KNN, GA, GA com KNN, GA com Média, Co-evolução, PSO
Classificadores	Número Nome do Classificador	2 Linear discriminant analysis, KNN	2 C4.5, CN2	2 New LERS, Naive LERS	2 Linear discriminant analysis, Logistic regression	6 RIPPER, C4.5, SVM (RBF Kernel), SVM (Polinomial Kernel), KNN, Naive Bayes	5 J48, IBK, Naive Bayes, PART, RBF Network
Bases de Dados		12	4	10	1	15	6
Qtd Valores Faltantes		1% a 20%	1% a 20%	1% a 13%	11%	5%, 10%, 20%, 30%, 40%, 50%	5%, 10%, 20%, 33%

4.1.1. Imputação utilizando Algoritmos Genéticos

Neste trabalho, propõe-se que o tratamento de valores faltantes em bases de dados pode ser modelado como um problema de busca/otimização, no qual se tenta maximizar o desempenho de um classificador encontrando os melhores valores a serem imputados. Nesse sentido, o uso de GA para realizar a imputação de valores faltantes se mostra uma estratégia promissora. Pode-se encontrar na literatura o uso de GA para tratar diversas tarefas de pré-processamento, como seleção de atributos, seleção de instâncias e construção de atributos (Freitas 2002). Porém, não são encontrados na literatura estudos abordando o uso sistemático de GA para fins de tratamento de valores faltantes, sendo o trabalho de (Abdella e Marwala 2005) uma exceção. Neste artigo, porém, o papel do GA é diferente, limitando-se a minimizar uma função de erro derivada a partir da aplicação de uma rede neural auto-associativa.

Foram utilizadas, neste trabalho, três variações de GA. Elas apresentam diferenças apenas em seus operadores de mutação, descritos a seguir:

- Mutação simples (*creep*): é feita apenas uma pequena alteração local ($\pm\delta$) em algum gene do indivíduo; este gene é escolhido de maneira aleatória;
- Mutação média/moda: nesta variação, atribui-se ao gene do indivíduo a média/moda do atributo que está recebendo a mutação; e
- Mutação KNN: nesta variação, atribui-se ao gene do indivíduo a média/moda dos *K* vizinhos mais próximos, estes encontrados utilizando o algoritmo KNN. Foi adotado K=3 neste trabalho.

Em todas as variações, a representação do indivíduo é linear. Cada valor faltante da base de dados é fixado em uma posição do indivíduo. Ou seja, caso a base possua *M* valores, porém apenas *N* sejam faltantes, o tamanho de cada indivíduo será *N*, conforme pode ser observado na Figura 8. No cálculo do valor de *fitness*, emprega-se uma estratégia *wrapper-like* (Kohavi e John 1997) (Freitas 2002) como discutido acima. O valor de *fitness* de cada indivíduo é dado pela taxa média de classificação correta (precisão) alcançada no processo de validação cruzada. O tamanho da população utilizado foi igual a 20, o número de gerações foi igual a 10, a taxa de *crossover* foi fixada em 70%, e a taxa de mutação em 10%.

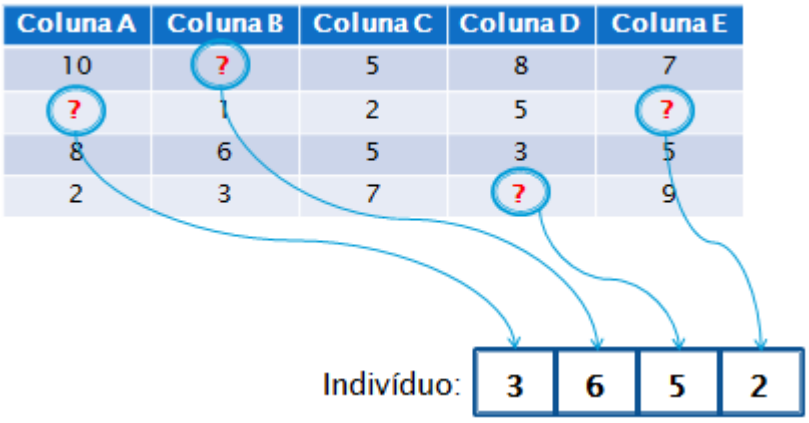

Figura 8: Representação do indivíduo para algoritmo genético

4.1.2. Imputação utilizando PSO

Até o presente, a utilização de PSO para tratamento de valores faltantes ainda não foi explorada na literatura. A decisão de utilizar o PSO foi motivada pelo bom desempenho que esse algoritmo vem demonstrando no tratamento de problemas de otimização numérica em geral (Kennedy, Eberhart e Shi 2001) (Clerc e Kennedy 2002) (de Castro 2007). Neste trabalho, foi utilizada a versão clássica do PSO, apresentada na Seção 3.2, com uma topologia do tipo global (todas as partículas são vizinhas entre si). A representação segue a mesma adotada para a imputação via GA, conforme Figura 9. O cálculo do *fitness* também é igual ao caso anterior, seguindo uma abordagem *wrapper-like*. O tamanho da população foi fixado em 20 partículas e o número de gerações foi igual a 10. Os valores dos demais parâmetros de controle estão definidos na Seção 3.2.

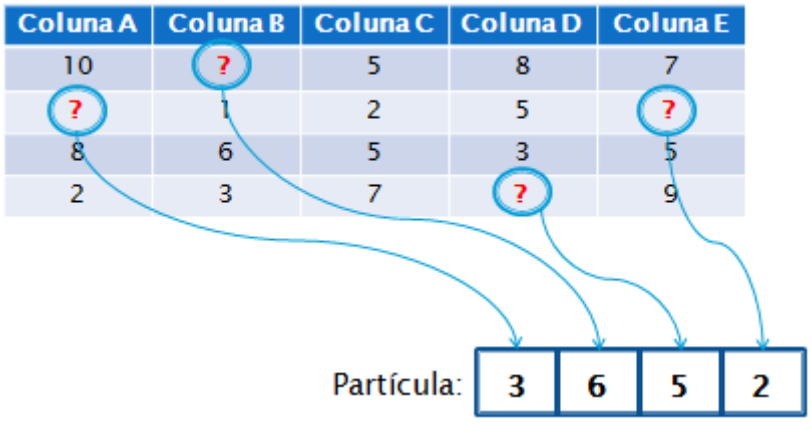

Figura 9: Representação da partícula para PSO

4.1.3. Imputação utilizando co-evolução

A última técnica bio-inspirada aplicada baseia-se na co-evolução cooperativa. A decisão de utilizar essa abordagem para o tratamento de valores faltantes foi motivada também pela inexistência de trabalhos na literatura, além do fato de ela seguir uma estratégia diferente de "divisão-e-conquista". As mesmas configurações do GA com mutação *creep* apresentadas anteriormente foram utilizadas no contexto da co-evolução. Foram utilizadas nos experimentos apenas duas espécies homogêneas, dividindo o tamanho do problema pela metade. Assim, se a base de dados possui *N* atributos faltantes, então o tamanho de cada indivíduo de uma espécie será *N/2*, conforme representado na Figura 10. A solução final de imputação é dada pela melhor combinação de indivíduos das duas espécies, a qual está armazenada na memória compartilhada (ver Seção 3.3).

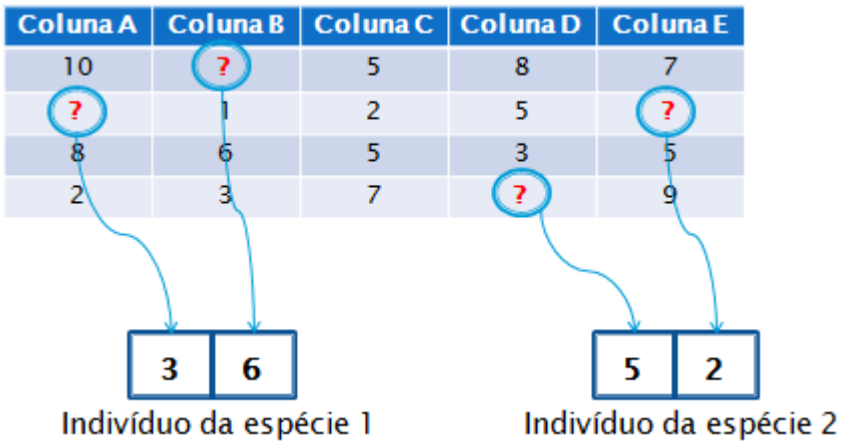

Figura 10: Representação do indivíduo para co-evolução

Para a avaliação do indivíduo no ambiente co-evolutivo foi empregada a abordagem com memória compartilhada. Foram guardadas sempre as três melhores combinações de indivíduos. O cálculo do valor de *fitness* é realizado também via validação cruzada de 10 pastas. O tamanho de cada população adotado nos experimentos foi de 10, metade do valor adotado para o GA simples. O número de gerações também foi 10, a taxa de *crossover* foi 70% e a taxa de mutação foi de 10%.

4.2. Resultados

Os resultados obtidos foram organizados em tabelas, que se encontram no Apêndice I. Foram divididos por algoritmo de classificação, taxa de atributos faltantes e técnica de imputação. Em cada tabela, as colunas *Crossval* e *Teste* denotam a média de precisão obtida pelo classificador utilizando a partição de validação cruzada e a de teste, respectivamente. Já a coluna *Wilcox* denota o *p*-valor obtido pela aplicação do teste estatístico não-paramétrico de Wilcoxon pareado[4] realizado entre o método em questão e a substituição pela média/moda. A coluna *Geração* indica, somente para as técnicas bio-inspiradas, a geração na qual o melhor indivíduo obtido ao longo do processo foi produzido pela primeira vez. Em destaque nas tabelas, observam-se os melhores resultados alcançados, em termos de média e desvio-padrão, tanto para validação cruzada como para teste.

4.2.1. Resultados para redes neurais RBF

Apresentaremos e discutiremos nesta subseção os resultados obtidos com redes neurais RBF. Na Figura 11 e na Figura 12, são dados os resultados alcançados em termos de validação cruzada. A primeira figura é relativa aos resultados obtidos com as bases *Iris*, *Zoo* e *Diabetes*, e a segunda é relativa às bases *BCD699*, *Car* e *Dermatology*. Por estes resultados, pode-se notar o bom desempenho das técnicas bio-inspiradas. Com exceção de um experimento, em todos os outros, as técnicas bio-inspiradas encontraram o melhor conjunto de dados a serem imputados para esse classificador.

A técnica utilizando co-evolução cooperativa apresentou resultados relevantes, mostrando ser a mais eficaz na validação cruzada em diversos experimentos. Pode ser apontada como uma boa alternativa de imputação de valores faltantes para este classificador. As técnicas utilizando variações do GA, aplicando KNN e média para mutação, saíram-se melhores do que o GA simples, apesar de as diferenças terem sido pequenas. Os três algoritmos encontraram bons resultados em termos de generalização sobre os dados de validação cruzada, geralmente acima dos obtidos pelas técnicas tradicionais. Já o PSO apresentou resultados bem variados, alcançando alguns

[4] O teste de Wilcoxon é o equivalente não-paramétrico do teste-t pareado (Hollander e Wolfe 1999). Nos experimentos, adotou-se o software R e α=5% como nível de significância, o que corresponde a um nível de confiança de 95%.

resultados relevantes, porém com o desempenho equivalente às técnicas tradicionais na maioria dos experimentos.

Analisando apenas os resultados obtidos nos testes (vide Apêndice I), nota-se uma grande proximidade entre os valores alcançados. Aplicando o teste não-paramétrico, percebe-se que a diferença entre a maioria dos resultados não é significante. As melhores taxas variaram entre as técnicas tradicionais e as bio-inspiradas. Considerando simultaneamente os resultados obtidos sobre as partições de validação cruzada e teste, pode-se afirmar que, em geral, as técnicas de imputação bio-inspiradas foram as mais robustas, ou seja, apresentaram taxas de precisão mais equivalentes. Nesse sentido, afirmamos que elas prevaleceram em termos de eficácia.

Na Tabela 5 são apresentados os tempos computacionais de pré-processamento exigidos para cada uma das técnicas. Evidencia-se o problema de escalabilidade enfrentado pela abordagem de imputação por otimização, principalmente para a técnica que utiliza co-evolução. Quanto maior a taxa de valores faltantes, maior a diferença no tempo de processamento entre as técnicas tradicionais e as bio-inspiradas. Nota-se também que, para a técnica que utiliza GA, as modificações na mutação não afetaram o tempo de processamento.

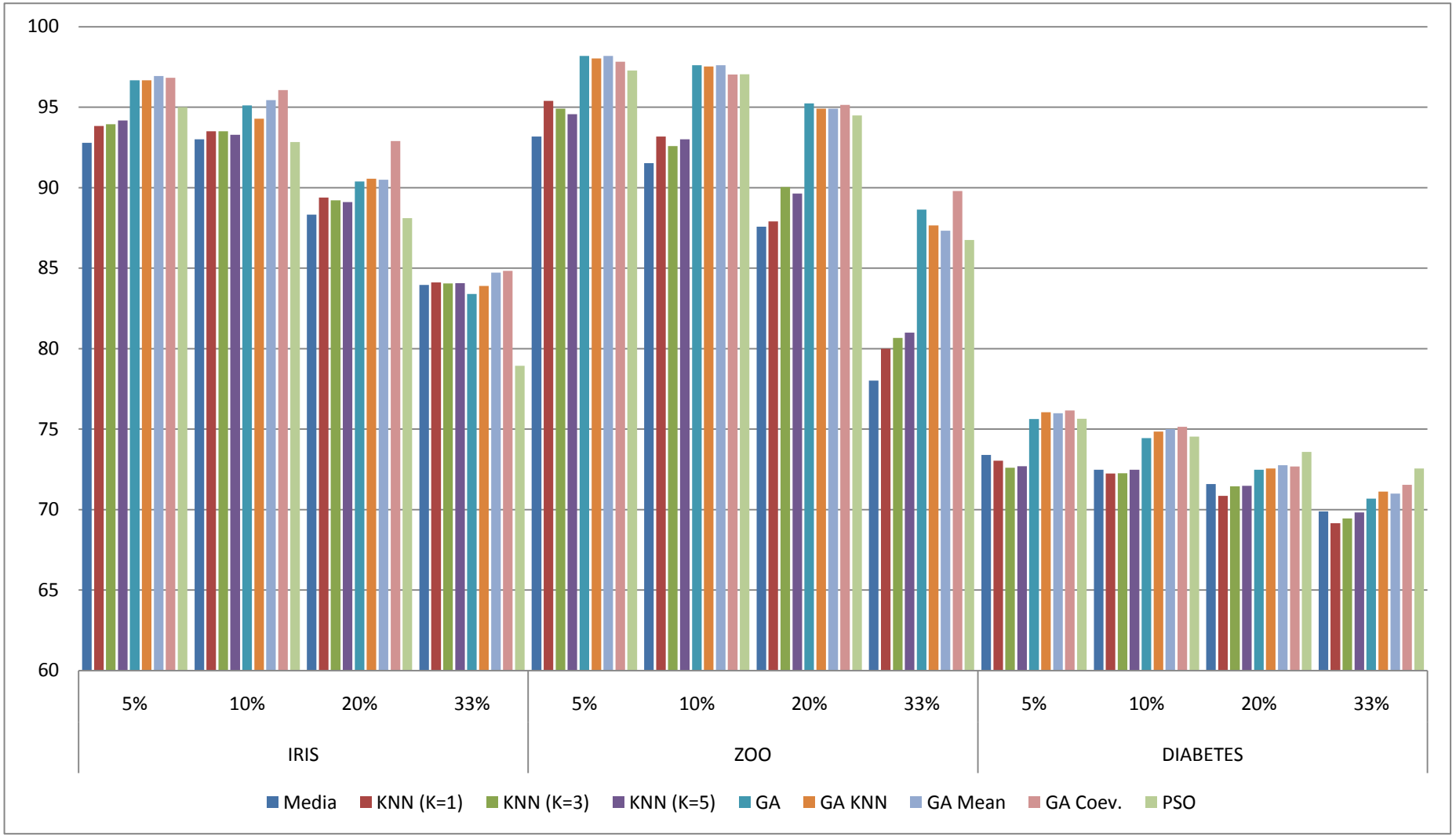

Figura 11: Resultados obtidos com redes neurais RBF – Iris, Zoo e Diabetes (validação cruzada)

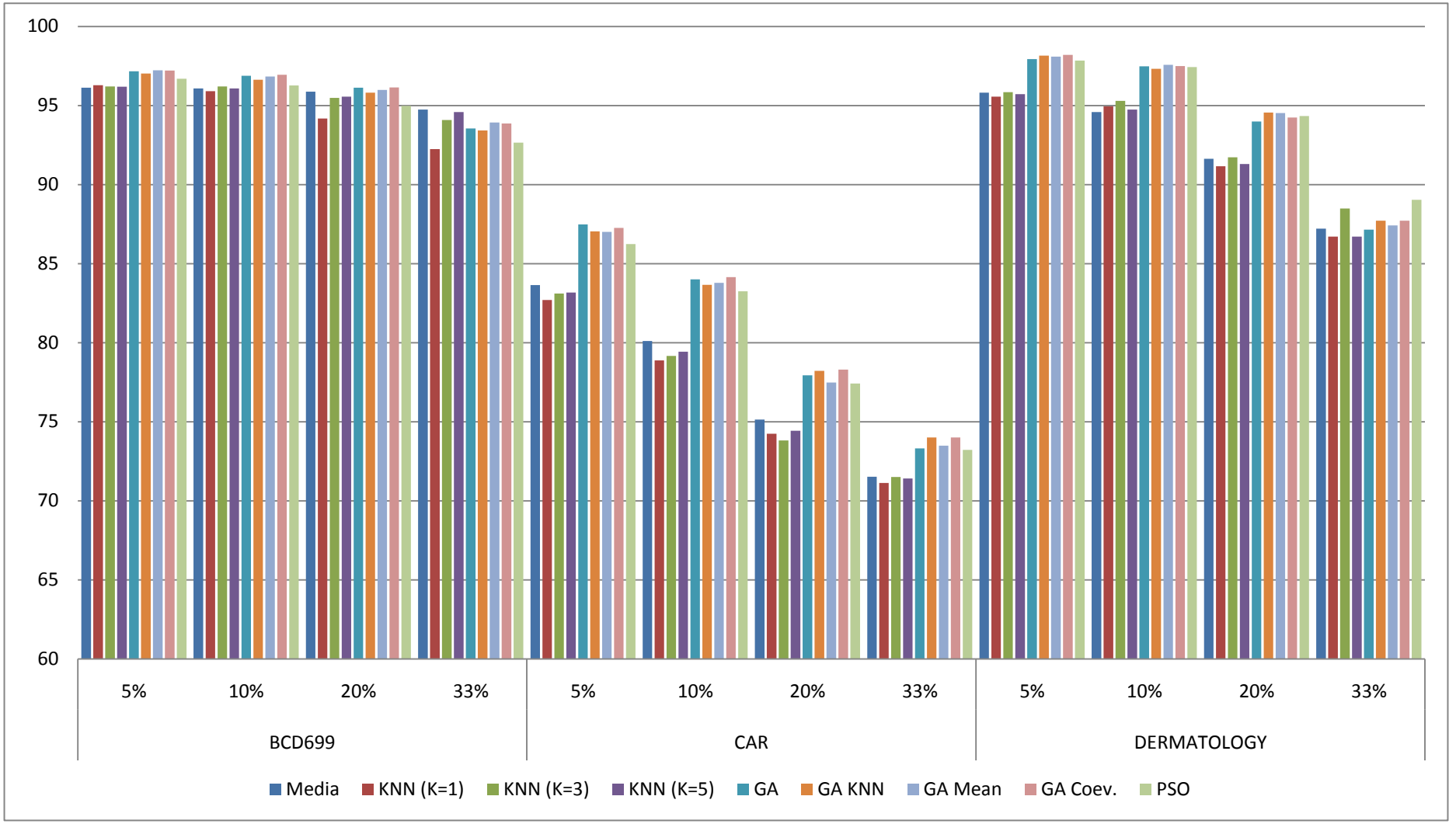

Figura 12: Resultados obtidos com redes neurais RBF – BCD699, Car e Dermatology (validação cruzada)

Tabela 5: Tempo de pré-processamento para redes neurais RBF

	5%	10%	20%	33%	5%	10%	20%	33%
Método	**Tempo de Execução (segundos)**							
	IRIS				**BCD699**			
Média	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,13±0,23	1,00±0,00	2,00±0,00	4,60±0,53
KNN (K=1)	0,00±0,00	0,00±0,00	0,07±0,12	0,00±0,00	1,07±0,12	1,00±0,00	2,00±0,00	2,40±0,69
KNN (K=3)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	2,00±0,00
KNN (K=5)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	2,00±0,00
GA	96,73±4,40	123,53±1,62	104,20±1,56	75,47±4,08	192,13±10,64	199,40±2,95	207,00±1,40	241,60±1,74
GA KNN	95,20±7,75	120,80±3,30	105,27±3,25	82,13±0,58	196,53±13,17	201,73±1,01	209,80±1,25	246,87±9,49
GA Mean	100,60±9,85	125,67±5,00	108,93±0,23	73,73±2,32	190,80±12,68	198,53±2,48	209,93±0,81	245,53±1,21
GA Coev.	161,27±6,49	238,47±13,34	245,80±25,17	184,40±7,30	546,73±20,33	643,13±31,82	977,40±45,18	1.750,53±23,84
PSO	100,47±8,39	120,93±1,68	96,13±1,33	70,67±0,70	208,80±12,44	246,87±14,77	331,27±6,31	478,33±0,76
	ZOO				**CAR**			
Média	0,00±0,00	0,20±0,20	0,73±0,12	1,47±0,12	6,20±0,20	7,40±0,35	9,47±0,46	12,73±0,64
KNN (K=1)	0,07±0,12	0,20±0,35	1,20±0,35	5,87±3,44	5,00±0,00	5,13±0,23	6,00±0,00	6,27±0,12
KNN (K=3)	0,07±0,12	0,47±0,23	0,93±0,12	3,47±2,91	5,07±0,12	5,07±0,12	6,00±0,00	6,07±0,12
KNN (K=5)	0,13±0,12	0,53±0,12	1,00±0,20	2,40±0,80	5,13±0,12	5,20±0,20	6,00±0,00	6,33±0,23
GA	125,67±9,03	139,07±4,66	162,60±10,34	241,13±26,68	898,40±79,15	872,00±28,93	838,60±4,10	902,20±3,75
GA KNN	140,00±9,04	146,47±7,78	177,40±4,01	261,00±5,00	865,93±11,00	873,73±17,71	855,07±16,91	913,87±7,43
GA Mean	120,20±15,25	143,00±2,25	164,13±7,56	237,07±7,31	873,87±23,39	882,47±48,30	859,93±43,10	897,40±4,57
GA Coev.	161,93±26,63	280,93±16,03	405,87±56,14	444,40±13,90	2.429,00±268,49	2.802,33±198,77	3.280,20±417,55	4.154,40±213,96
PSO	150,07±9,93	157,20±10,54	184,13±30,81	217,40±2,42	951,13±47,26	1.064,00±50,54	1.252,20±6,73	1.540,40±15,07
	DIABETES				**DERMATOLOGY**			
Média	1,00±0,00	1,00±0,00	2,20±0,20	5,33±0,23	4,53±1,42	7,07±2,61	25,47±9,54	75,73±17,97
KNN (K=1)	1,00±0,00	1,00±0,00	2,00±0,00	2,40±0,53	5,80±1,83	8,47±1,45	24,33±4,20	35,40±3,56
KNN (K=3)	1,00±0,00	1,00±0,00	2,00±0,00	2,27±0,31	5,20±1,22	6,93±0,76	23,33±1,14	43,40±5,05
KNN (K=5)	1,00±0,00	1,00±0,00	2,00±0,00	2,13±0,23	5,53±1,63	8,20±1,25	22,93±3,97	52,60±13,44
GA	181,27±4,05	192,67±1,86	210,53±0,70	249,67±5,20	596,27±183,15	734,67±120,89	1.702,47±395,70	4.006,27±689,68
GA KNN	189,80±3,82	199,40±3,80	218,87±0,76	248,07±2,00	608,93±176,32	865,07±170,13	2.329,80±730,89	5.611,80±235,44
GA Mean	179,27±2,81	190,27±2,42	207,67±1,42	238,67±2,23	570,87±214,54	853,33±151,07	1.774,07±387,80	4.560,87±470,29
GA Coev.	569,00±1,59	699,40±21,12	1.037,47±12,02	1.746,93±37,33	1.505,60±204,58	2.076,20±438,02	4.320,33±1.130,14	10.561,07±570,75
PSO	207,93±2,91	254,80±2,62	355,53±2,87	493,80±6,16	642,00±206,31	987,00±393,48	2.258,00±516,07	5.714,47±744,61

4.2.2. Resultados para o algoritmo J48

Os resultados obtidos com o algoritmo J48 serão apresentados nesta subseção. Na Figura 13 e Figura 14, os resultados obtidos na fase de validação cruzada são apresentados graficamente. Novamente, pode-se observar o bom desempenho das técnicas bio-inspiradas nesta fase, as quais, em geral, obtiveram resultados superiores aos das técnicas tradicionais.

Para este classificador, os resultados em sua maioria foram bem competitivos entre as técnicas bio-inspiradas. Contudo, nota-se a técnica utilizando PSO se sobressaindo em diversos experimentos, tanto no que tange à validação cruzada como à etapa de teste (vide Apêndice I). Para a base *Dermatology*, que é a mais complexa das bases, a técnica utilizando PSO alcançou resultados bem acima dos obtidos pelas outras técnicas, inclusive as bio-inspiradas, tanto na validação cruzada como no teste, mostrando-se ser uma boa alternativa de técnica de imputação para esse classificador.

Analisando as técnicas que utilizam GA, incluindo a co-evolutiva, nota-se uma grande competitividade entre elas. Porém, a técnica co-evolutiva se destaca, alcançando melhores resultados na maioria dos experimentos. Comparando apenas a técnica utilizando GA simples e suas variações, com KNN e média, também se percebe que, para o algoritmo J48, o GA não foi muito sensível a essas variações. Para a base *Dermatology*, o GA se mostrou muito ineficaz, obtendo resultados abaixo dos alcançados pelas técnicas tradicionais.

Na Tabela 6 são apresentados os tempos de pré-processamento para os experimentos realizados com o algoritmo J48. Novamente se nota a dificuldade das técnicas bio-inspiradas em relação ao quesito escalabilidade. As técnicas utilizando PSO e co-evolução foram as de pior desempenho nesse sentido. Enquanto os métodos tradicionais executam em menos de 10 segundos, o PSO, por exemplo, chegou às vezes a consumir mais de 90 minutos para finalizar o processamento. Assim como para as redes neurais RBF, não houve grandes variações em relação ao tempo de processamento quando comparamos as três variações de GA utilizadas.

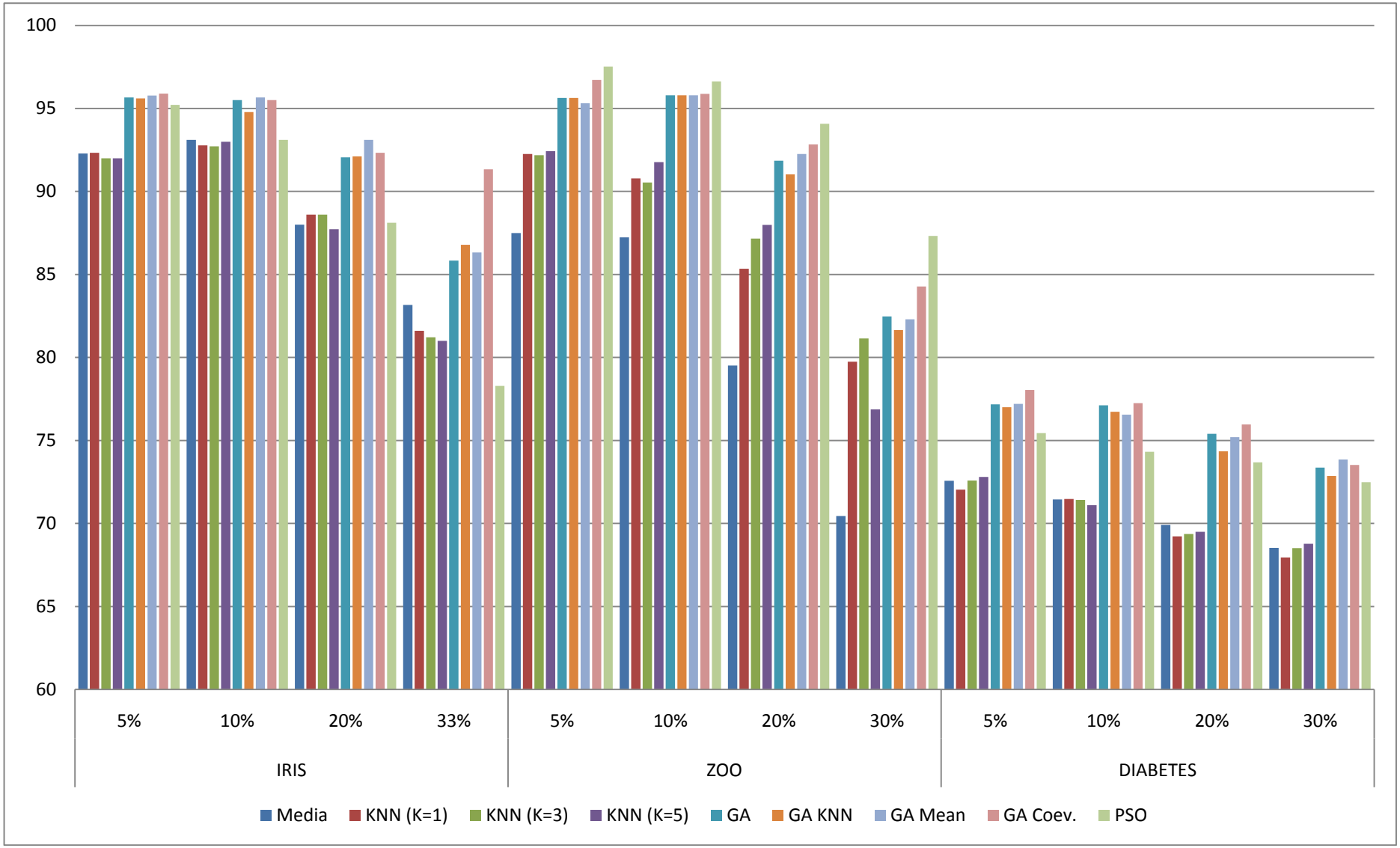

Figura 13: Resultados obtidos para o algoritmo J48 – Iris, Zoo e Diabetes (validação cruzada)

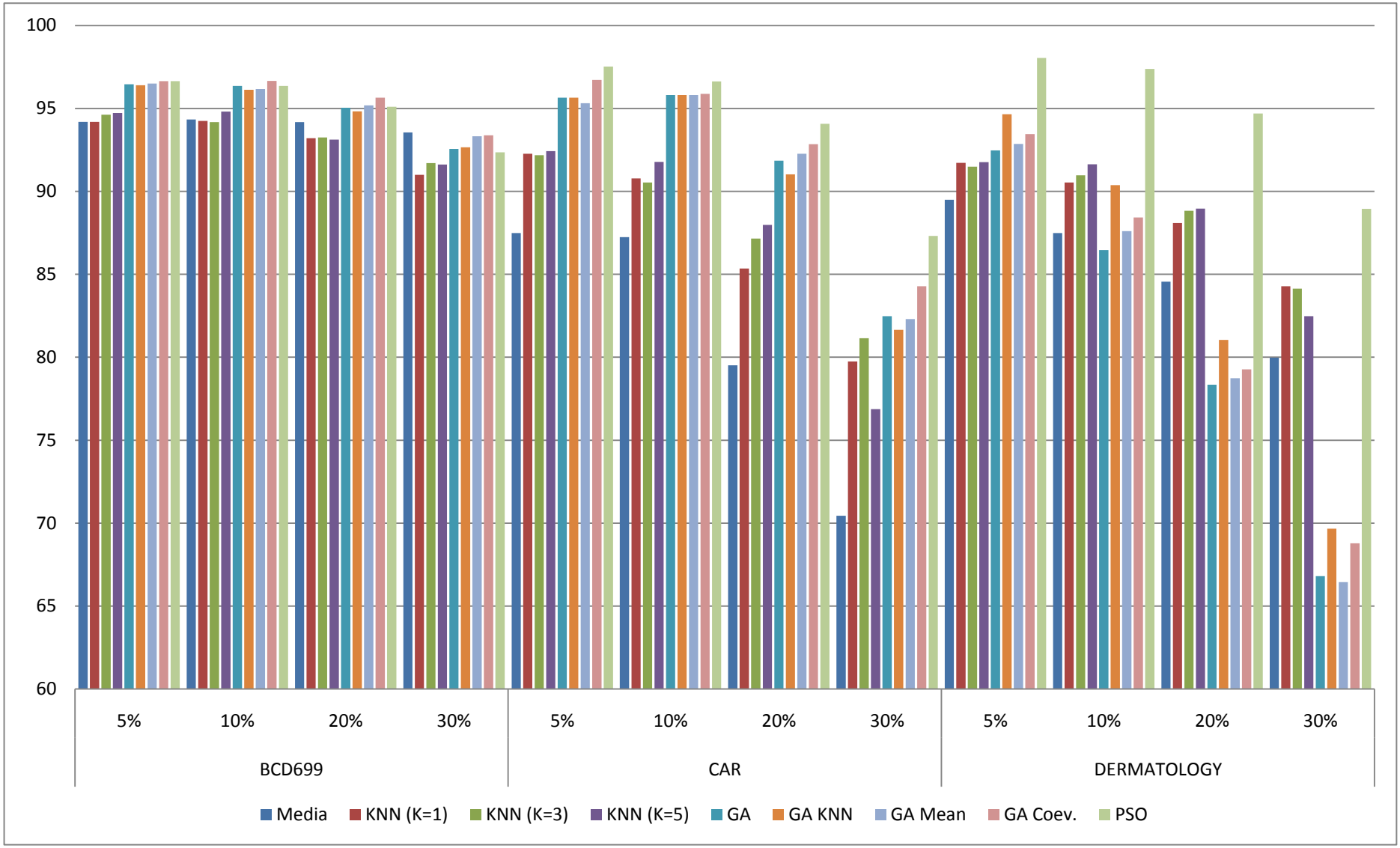

Figura 14: Resultados obtidos para o algoritmo J48 – BCD699, Car e Dermatology (validação cruzada)

Tabela 6: Tempo de pré-processamento para o algoritmo J48

	5%	10%	20%	33%	5%	10%	20%	33%
Método	**Tempo de Execução (segundos)**							
	IRIS				**BCD699**			
Média	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	0,47±0,81	0,67±0,58	1,33±0,42	3,40±0,53
KNN (K=1)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	0,40±0,35	1,00±0,00	1,00±0,00	2,00±0,00
KNN (K=3)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00
KNN (K=5)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	0,13±0,23	1,00±0,00	1,00±0,00	2,00±0,00
GA	35,40±0,72	36,20±1,11	38,07±0,61	39,13±1,10	66,47±1,89	72,20±1,64	91,67±2,00	130,33±2,21
GA KNN	37,27±1,30	36,87±1,21	38,27±1,14	39,87±0,81	66,87±0,50	73,20±1,40	91,93±1,81	128,00±1,22
GA Mean	38,67±5,90	40,80±6,66	42,07±6,22	41,00±1,00	65,47±0,64	71,33±0,83	90,33±2,02	128,80±2,75
GA Coev.	66,67±9,45	89,40±4,26	104,67±4,77	114,13±3,16	293,60±12,90	397,67±6,71	647,80±16,25	1.336,67±214,50
PSO	102,33±8,60	123,47±1,17	99,00±2,78	75,47±5,56	209,47±20,26	234,40±6,42	319,40±12,70	454,60±22,33
	ZOO				**CAR**			
Média	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	2,00±0,00	3,07±0,12	5,47±0,50	8,40±0,35
KNN (K=1)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	1,93±0,12	2,00±0,00	3,00±0,00
KNN (K=3)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	1,80±0,20	2,00±0,00	3,00±0,00
KNN (K=5)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,00±0,00	2,20±0,35	2,67±1,15	3,67±1,15
GA	35,47±0,58	36,27±1,03	37,53±1,01	39,00±0,80	83,27±3,82	100,87±3,93	143,20±1,78	206,27±1,80
GA KNN	60,47±0,50	60,53±1,29	61,80±1,20	63,87±1,01	88,87±4,04	106,93±3,56	148,20±2,31	210,87±1,89
GA Mean	36,80±0,69	37,93±1,01	39,20±1,20	41,27±1,42	84,20±6,68	99,80±4,33	141,40±3,98	205,20±4,89
GA Coev.	63,80±1,25	93,93±5,15	110,20±2,23	119,00±3,42	864,47±39,12	1.127,07±26,88	1.533,07±46,96	2.173,53±21,77
PSO	141,67±10,82	151,53±15,50	173,73±19,17	232,67±25,80	960,80±61,73	1.082,20±68,62	1.271,07±8,93	1.512,13±8,31
	DIABETES				**DERMATOLOGY**			
Média	0,80±0,35	1,00±0,00	2,00±0,00	4,00±0,00	0,73±0,12	1,00±0,00	2,00±0,00	3,47±0,50
KNN (K=1)	1,00±0,00	1,00±0,00	1,07±0,12	2,00±0,00	0,80±0,20	1,00±0,00	2,00±0,00	2,80±0,20
KNN (K=3)	1,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	0,67±0,12	1,00±0,00	2,00±0,00	2,87±0,23
KNN (K=5)	1,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	0,73±0,31	1,00±0,00	2,00±0,00	2,87±0,12
GA	92,20±5,84	96,60±7,54	112,20±7,98	141,73±8,55	67,53±0,99	83,40±1,04	123,67±0,70	193,33±0,81
GA KNN	88,33±2,84	95,47±1,33	109,60±1,40	137,47±1,67	69,27±1,10	85,40±0,87	126,13±0,99	195,13±0,64
GA Mean	91,53±15,81	93,47±7,56	104,40±2,60	135,20±1,91	66,20±1,22	82,00±0,92	121,87±1,10	190,93±0,46
GA Coev.	461,53±35,20	597,40±39,36	916,87±9,38	1.819,67±408,79	386,40±15,59	547,53±21,27	862,53±25,76	1.508,40±227,77
PSO	207,80±2,00	255,80±1,56	355,07±3,30	490,20±2,96	617,87±191,31	1.010,93±374,52	2.256,93±925,96	5.403,33±826,97

4.2.3. Resultados para o algoritmo IBK

Nesta subseção, apresentaremos e discutiremos os resultados obtidos com o algoritmo IBK. Na Figura 15 e na Figura 16, temos os resultados alcançados, em termos de eficácia sobre as partições de validação cruzada, para esse classificador. A primeira figura é relativa aos resultados obtidos com as bases *Iris*, *Zoo* e *Diabetes*, e a segunda, relativa às bases *BCD699*, *Car* e *Dermatology*. Os resultados se mostraram bem competitivos, apresentando variações bem sutis entre os métodos utilizados. Novamente, as técnicas bio-inspiradas se mostraram mais eficazes sobre a partição de validação cruzada, com os melhores resultados variando entre as técnicas que utilizam GA e a que utiliza PSO.

O destaque fica com a técnica que utiliza PSO, que obteve bons resultados nas bases mais complexas, tanto na validação cruzada como na etapa de teste (vide Apêndice I). Mostrou ser uma boa alternativa para este classificador, já que, para as bases mais simples, obteve resultados competitivos. As variações de GA utilizando a média e o KNN na mutação apresentaram resultados bem próximos, na maioria das vezes melhor que o GA simples. A técnica utilizando co-evolução também apresentou resultados próximos aos dos GA modificados.

Observa-se também que, para os testes (vide Apêndice I), os resultados foram bem variados. As técnicas tradicionais alcançaram os melhores resultados em alguns experimentos. Pode-se observar ainda que poucas vezes uma técnica foi a mais eficaz na validação cruzada e alcançou o melhor resultado também no teste. Este fato sugere que, apesar de obterem um bom modelo para os dados de validação cruzada, os modelos não eram genéricos o suficiente para classificar as amostras das partições de teste.

A Tabela 7 contém os tempos de execução dos experimentos para o algoritmo IBK. Os valores apontam que, assim como para os classificadores anteriores, as técnicas bio-inspiradas apresentaram os maiores valores. As técnicas que utilizam o PSO e a co-evolução sempre apresentam um pior desempenho nesse quesito.

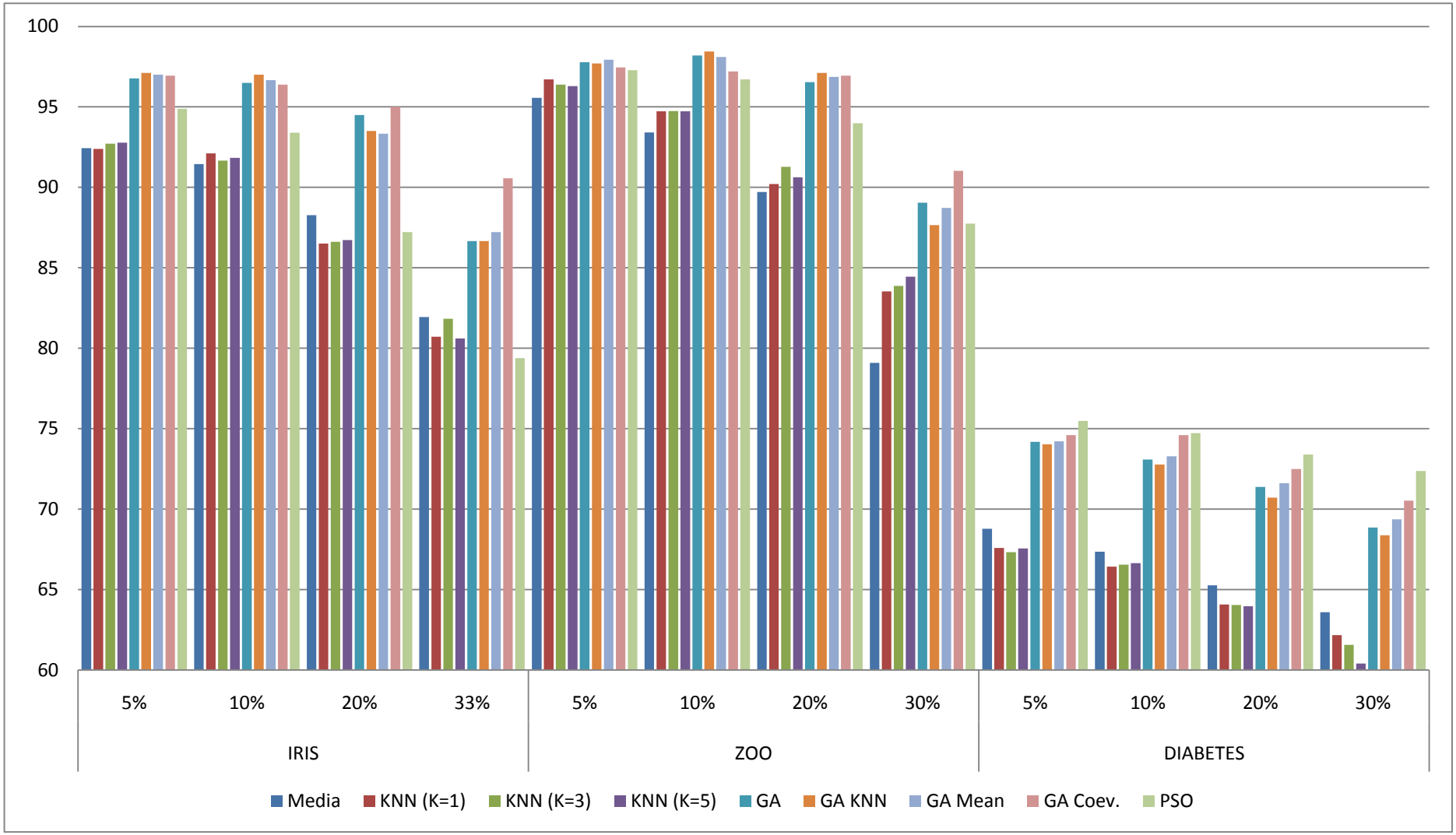

Figura 15: Resultados obtidos com o algoritmo IBK – Iris, Zoo e Diabetes (validação cruzada)

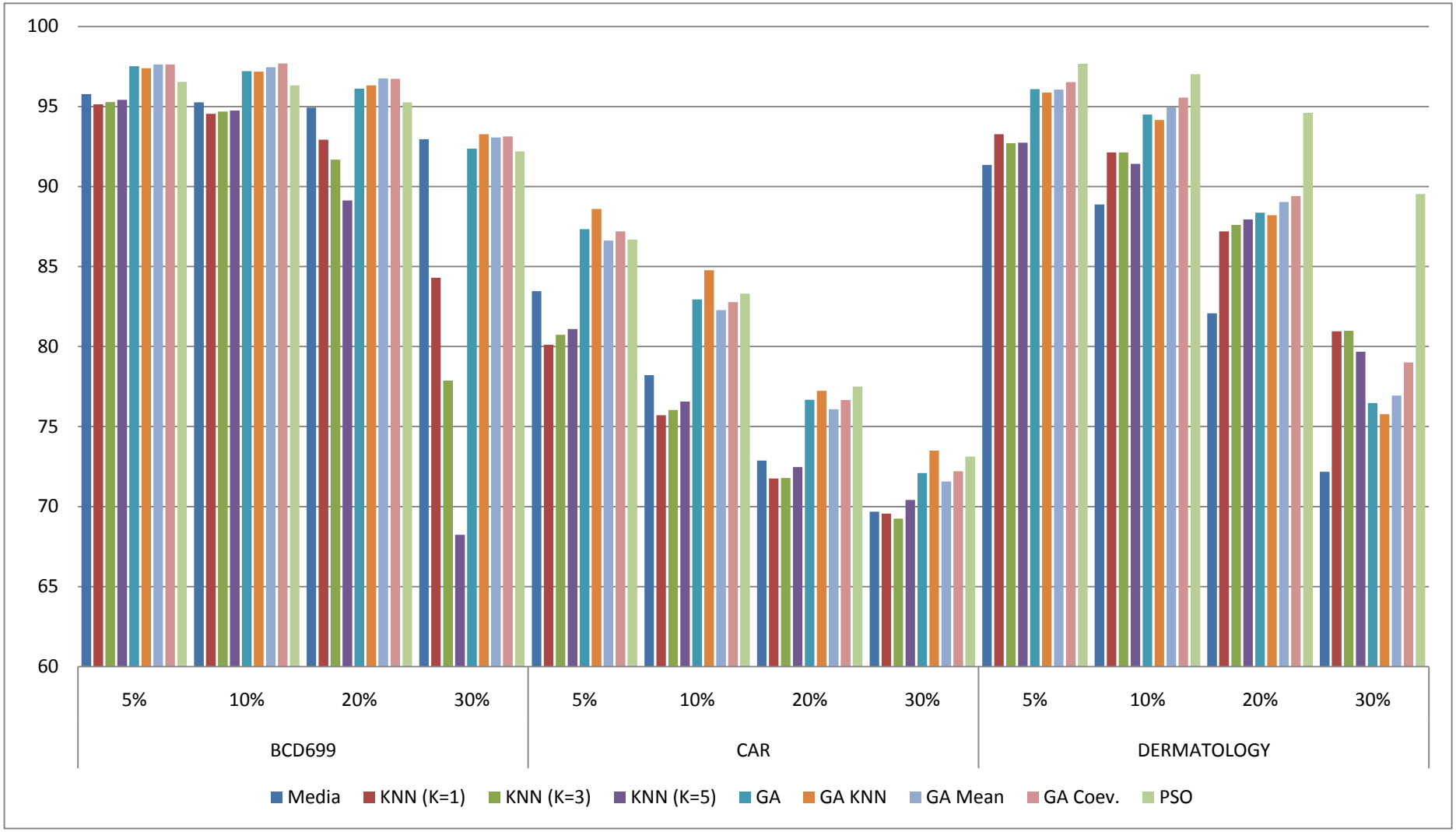

Figura 16: Resultados obtidos com o algoritmo IBK – BCD699, Car e Dermatology (validação cruzada)

Tabela 7: Tempo de pré-processamento para o algoritmo IBK

	5%	10%	20%	33%	5%	10%	20%	33%
Método	**Tempo de Execução (segundos)**							
	IRIS				**BCD699**			
Média	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,47±0,81	1,47±0,50	2,33±0,42	4,27±0,31
KNN (K=1)	0,00±0,00	0,07±0,12	0,00±0,00	0,00±0,00	1,20±0,20	1,87±0,12	2,00±0,00	3,00±0,00
KNN (K=3)	0,00±0,00	0,07±0,12	0,00±0,00	0,00±0,00	1,07±0,12	1,67±0,12	2,00±0,00	3,00±0,00
KNN (K=5)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	1,13±0,23	1,67±0,31	2,00±0,00	2,93±0,12
GA	40,93±4,23	42,87±5,00	43,60±5,72	41,80±2,62	120,27±8,27	126,13±8,73	141,93±10,81	164,27±0,64
GA KNN	38,60±1,44	38,93±1,10	39,27±0,70	40,13±1,01	118,80±1,83	123,33±0,58	136,73±0,64	164,67±0,81
GA Mean	37,80±1,31	38,93±1,10	39,73±1,10	40,00±1,00	126,33±1,01	132,27±0,76	147,13±0,70	176,00±0,35
GA Coev.	64,13±3,72	93,00±6,16	106,80±9,30	109,60±3,83	396,33±5,83	469,13±14,14	722,60±27,27	1.376,40±193,11
PSO	121,53±9,40	136,20±1,11	113,47±7,92	87,00±9,77	198,87±3,60	226,13±4,39	301,47±1,89	447,60±6,24
	ZOO				**CAR**			
Média	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	4,67±0,70	5,87±0,83	8,40±0,72	11,13±0,58
KNN (K=1)	0,07±0,12	0,00±0,00	0,00±0,00	0,07±0,12	3,07±0,12	3,93±0,12	4,07±0,12	5,07±0,12
KNN (K=3)	0,00±0,00	0,00±0,00	0,00±0,00	0,00±0,00	3,07±0,12	4,00±0,00	4,13±0,12	5,07±0,12
KNN (K=5)	0,13±0,23	0,07±0,12	0,00±0,00	0,00±0,00	3,07±0,12	4,00±0,20	4,47±0,46	5,27±0,46
GA	37,67±3,38	40,00±0,53	40,73±1,42	42,13±1,03	315,00±1,74	330,67±1,70	369,33±1,10	431,33±0,42
GA KNN	62,93±1,10	58,53±5,62	62,47±3,56	65,80±1,31	326,00±1,91	341,67±1,33	381,00±0,87	445,33±0,81
GA Mean	39,00±1,80	41,87±1,03	42,07±1,10	43,47±1,33	319,87±5,66	334,20±4,69	374,80±6,77	437,13±3,97
GA Coev.	62,47±1,68	92,47±2,84	108,60±6,88	115,53±1,70	1.329,20±55,01	1.534,93±19,80	2.021,67±8,17	2.597,13±70,17
PSO	135,00±5,57	138,60±2,16	161,20±2,82	213,27±2,21	983,73±75,14	1.114,07±89,45	1.301,67±40,00	1.533,80±5,05
	DIABETES				**DERMATOLOGY**			
Média	1,07±0,12	2,00±0,00	3,00±0,00	5,00±0,00	1,00±0,00	2,00±0,00	3,00±0,00	4,00±0,00
KNN (K=1)	1,73±0,46	2,00±0,00	2,07±0,12	3,00±0,00	1,00±0,00	2,00±0,00	2,27±0,12	3,13±0,23
KNN (K=3)	1,67±0,23	2,00±0,00	2,00±0,00	3,00±0,00	1,07±0,12	2,00±0,00	2,20±0,20	3,07±0,12
KNN (K=5)	1,87±0,23	2,00±0,00	2,07±0,12	3,00±0,00	1,00±0,00	2,00±0,00	2,07±0,12	3,00±0,00
GA	128,13±3,25	133,80±0,72	151,13±0,12	182,73±0,81	103,47±1,33	117,87±1,10	155,00±1,00	222,00±1,91
GA KNN	127,67±2,08	133,20±0,20	150,13±0,90	182,00±0,40	115,27±20,05	120,33±4,74	155,13±0,58	221,87±1,17
GA Mean	124,00±1,00	131,20±1,06	149,20±0,80	180,80±0,69	103,73±0,81	117,47±0,42	155,07±0,90	221,80±0,72
GA Coev.	522,20±5,76	607,40±17,29	956,07±34,01	1.841,13±382,91	471,47±22,50	623,47±4,00	924,07±42,95	1.570,60±146,20
PSO	212,13±4,24	262,00±1,91	361,00±2,80	499,47±6,00	660,00±248,30	956,13±198,54	2.386,33±463,15	4.790,33±794,66

4.2.4. Resultados para o algoritmo Naïve Bayes

Nesta subseção, serão apresentados e discutidos os resultados obtidos com o classificador *Naïve Bayes* sobre os dados de validação cruzada. Os resultados são ilustrados na Figura 17 e na Figura 18. Nota-se novamente a superioridades das técnicas bio-inspiradas frente às abordagens tradicionais. Contudo, assim como para o algoritmo IBK, os resultados para este classificador foram bem competitivos. As técnicas bio-inspiradas alcançaram resultados bem próximos das técnicas tradicionais.

O destaque aqui fica por conta da técnica co-evolutiva, que alcançou bons resultados tanto nas bases simples como nas mais complexas. Já técnica utilizando PSO, que vinha se destacando com os classificadores anteriores, não apresentou um bom desempenho para o *Naïve Bayes*. Os resultados obtidos com as técnicas utilizando GA e suas variações foram, em geral, bem próximos, não apresentando diferenças consideráveis. Contudo, analisando a geração do melhor indivíduo, percebe-se que a técnica que utiliza GA modificado com KNN consegue convergir mais rapidamente que as demais. A técnica co-evolutiva também conseguiu, em alguns experimentos, convergir com um baixo número de gerações. Contudo, a técnica utilizando PSO é a que necessitou em geral do menor número de gerações para alcançar o (quase-)ótimo.

Pode-se observar também grandes variações nos resultados obtidos para as partições de teste (vide Apêndice I), sendo que os melhores índices de desempenho em termos de precisão se alternam entre as técnicas tradicionais e bio-inspiradas. Novamente, dificilmente a técnica mais eficaz na etapa de validação cruzada obteve o melhor valor na fase de teste. Entretanto, os valores encontrados foram bem próximos, o que sugere um nível razoável de generalização do modelo induzido.

Com relação ao tempo de processamento, apresentados na Tabela 8, o problema de escalabilidade das técnicas bio-inspiradas fica mais uma vez evidenciado. Novamente, as técnicas co-evolutiva e com PSO apresentam os piores valores nesse quesito.

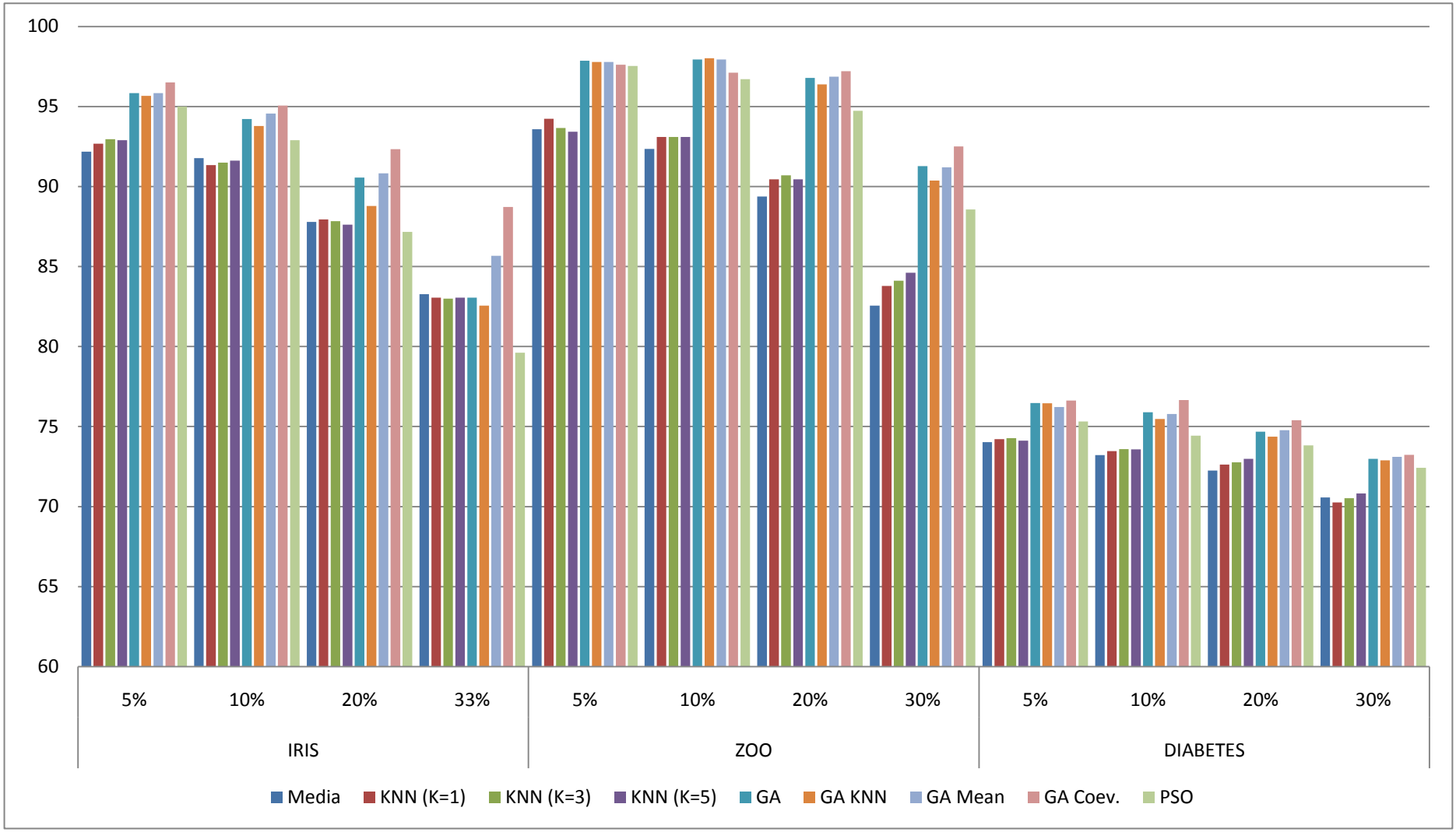

Figura 17: Resultados obtidos com o algoritmo Naïve Bayes – Iris, Zoo e Diabetes (validação cruzada)

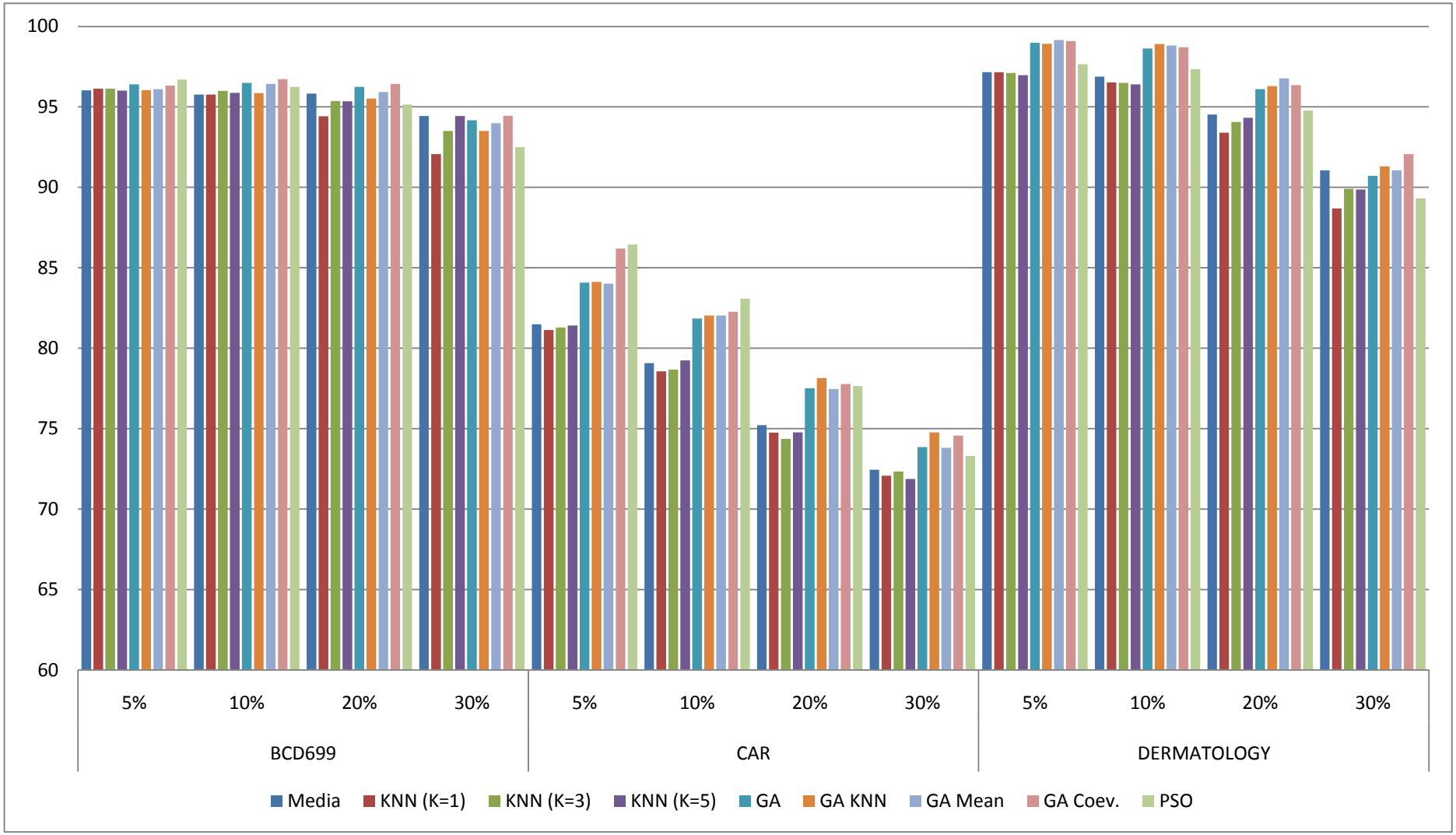

Figura 18: Resultados obtidos com o algoritmo Naïve Bayes – BCD699, Car e Dermatology (validação cruzada)

Tabela 8: Tempo de pré-processamento para o algoritmo Naïve Bayes

	5%	10%	20%	33%	5%	10%	20%	33%
Método	**Tempo de Execução**							
	IRIS				**BCD699**			
Média	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,27±0,64	2,07±0,12	2,87±0,42	4,87±0,23
KNN (K=1)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,13±0,23	2,00±0,00	2,40±0,40	3,00±0,00
KNN (K=3)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	2,00±0,00	2,53±0,31	3,00±0,00
KNN (K=5)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,00±0,00	2,00±0,00	2,20±0,20	3,00±0,00
GA	39,40±0,72	40,00±1,00	41,40±1,51	42,00±1,06	47,07±2,08	51,13±0,76	65,67±1,03	94,73±0,50
GA KNN	40,73±1,27	40,73±0,70	41,33±0,58	41,80±0,80	50,13±1,03	55,33±1,17	69,53±0,99	97,20±0,69
GA Mean	39,93±0,61	40,60±0,53	41,07±0,90	41,73±0,64	51,20±0,80	56,40±0,53	71,40±1,06	99,53±0,70
GA Coev.	55,87±1,70	77,93±4,82	100,07±4,02	108,53±3,25	262,07±12,53	341,07±15,83	561,87±16,74	1.241,20±235,32
PSO	118,00±12,28	130,40±2,09	104,00±3,65	81,07±3,59	206,67±2,32	239,20±6,35	320,07±1,10	462,33±2,91
	ZOO				**CAR**			
Média	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	5,27±0,83	6,47±0,64	9,07±1,01	11,93±0,76
KNN (K=1)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	4,00±0,00	4,07±0,12	5,00±0,00	5,40±0,20
KNN (K=3)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	4,00±0,00	4,13±0,23	5,00±0,00	5,33±0,23
KNN (K=5)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	4,00±0,00	4,27±0,23	5,40±0,53	6,07±0,90
GA	36,47±4,62	41,07±0,90	41,73±0,64	43,00±0,72	56,40±1,51	71,27±1,22	110,07±0,64	172,53±0,12
GA KNN	59,07±5,25	63,67±3,53	66,00±1,00	67,40±0,87	66,60±2,09	81,07±1,01	118,60±1,25	182,80±1,60
GA Mean	38,40±4,23	41,73±1,72	41,87±1,22	44,27±1,27	57,47±0,64	72,93±0,23	111,40±0,72	175,53±0,83
GA Coev.	61,73±5,70	91,67±5,59	100,13±5,86	113,53±2,52	835,93±13,51	1.059,13±18,22	1.459,60±21,00	2.079,27±62,25
PSO	135,27±5,01	140,13±3,44	165,33±13,80	216,20±29,72	991,40±59,19	1.120,67±63,35	1.303,27±15,32	1.563,60±13,34
	DIABETES				**DERMATOLOGY**			
Média	2,00±0,00	2,00±0,00	3,07±0,12	5,67±0,31	2,00±0,00	2,00±0,00	3,00±0,00	4,93±0,12
KNN (K=1)	2,00±0,00	2,00±0,00	3,00±0,00	3,13±0,12	2,00±0,00	2,00±0,00	3,00±0,00	4,00±0,00
KNN (K=3)	2,00±0,00	2,13±0,12	3,00±0,00	3,13±0,12	2,00±0,00	2,00±0,00	3,00±0,00	4,00±0,00
KNN (K=5)	2,00±0,00	2,20±0,20	3,00±0,00	3,00±0,00	2,00±0,00	2,00±0,00	3,00±0,00	4,00±0,00
GA	51,93±0,90	58,80±0,35	76,53±0,70	108,60±1,22	55,13±1,01	67,07±4,09	106,07±0,90	173,27±1,86
GA KNN	52,80±1,04	59,67±1,01	77,20±1,40	107,73±0,31	56,73±0,61	70,53±0,50	108,27±0,12	174,20±1,44
GA Mean	53,00±0,92	59,47±1,36	76,93±1,14	107,13±1,22	54,00±3,56	69,47±0,81	107,87±1,40	173,33±0,42
GA Coev.	375,80±4,87	512,20±8,67	826,00±43,58	1.708,73±364,30	351,47±26,49	537,80±18,68	866,60±11,53	1.500,07±212,80
PSO	213,47±1,75	266,67±5,69	364,87±10,59	535,33±57,29	683,47±220,70	936,07±293,62	2.209,47±507,83	5.420,20±1.154,56

4.2.5. Resultados para o algoritmo PART

Nesta subseção, os resultados obtidos com o algoritmo PART são apresentados. Na Figura 19 e na Figura 20, temos os resultados obtidos em termos de taxa de classificação correta sobre as partições de validação cruzada. Novamente, notam-se resultados bastante competitivos, principalmente entre as técnicas bio-inspiradas.

A técnica co-evolutiva e a técnica com PSO foram, novamente, as mais eficazes. Para este classificador, pode-se notar uma grande alternância de melhor desempenho entre as duas, principalmente nas duas bases mais complexas. A técnica co-evolutiva foi, de longe, a mais eficaz na base *Car*, enquanto a técnica com PSO obteve resultados muito aquém. Entretanto, para a base *Dermatology*, a técnica utilizando PSO obteve resultados muito expressivos, ficando consideravelmente acima de todas as outras técnicas. Já a técnica com co-evolução não conseguiu superar os resultados obtidos pelas técnicas que utilizam GA, ficando abaixo até das técnicas tradicionais. As técnicas utilizando as três variações de GA alcançaram resultados bem próximos. As variações do GA não trouxeram grandes ganhos, uma vez que o GA simples conseguiu resultados superiores em parte dos experimentos.

Analisando as gerações em que os melhores indivíduos foram obtidos (que dá uma noção da qualidade de convergência dos métodos bio-inspirados), observa-se certa vantagem da técnica utilizando PSO sobre as outras. Considerando-se apenas as técnicas com GA, a co-evolução apresenta os menores valores, ficando mais próxima do PSO.

A variabilidade entre o melhor desempenho na etapa de validação cruzada e o melhor desempenho na etapa de teste se repetiu aqui (vide Apêndice I). Raramente, os melhores modelos gerados na validação cruzada alcançaram também melhores resultados nos testes. As exceções foram para os experimentos mais complexos (*Car* e *Dermatology* – 20% e 33%).

Em relação ao tempo de processamento, apresentados na Tabela 9, novamente pode se constatar o baixo desempenho das técnicas bio-inspiradas. Conforme os demais experimentos, as técnicas baseadas em co-evolução e PSO são as que necessitam de maior tempo de pré-processamento, ficando bem acima do tempo das outras técnicas utilizadas.

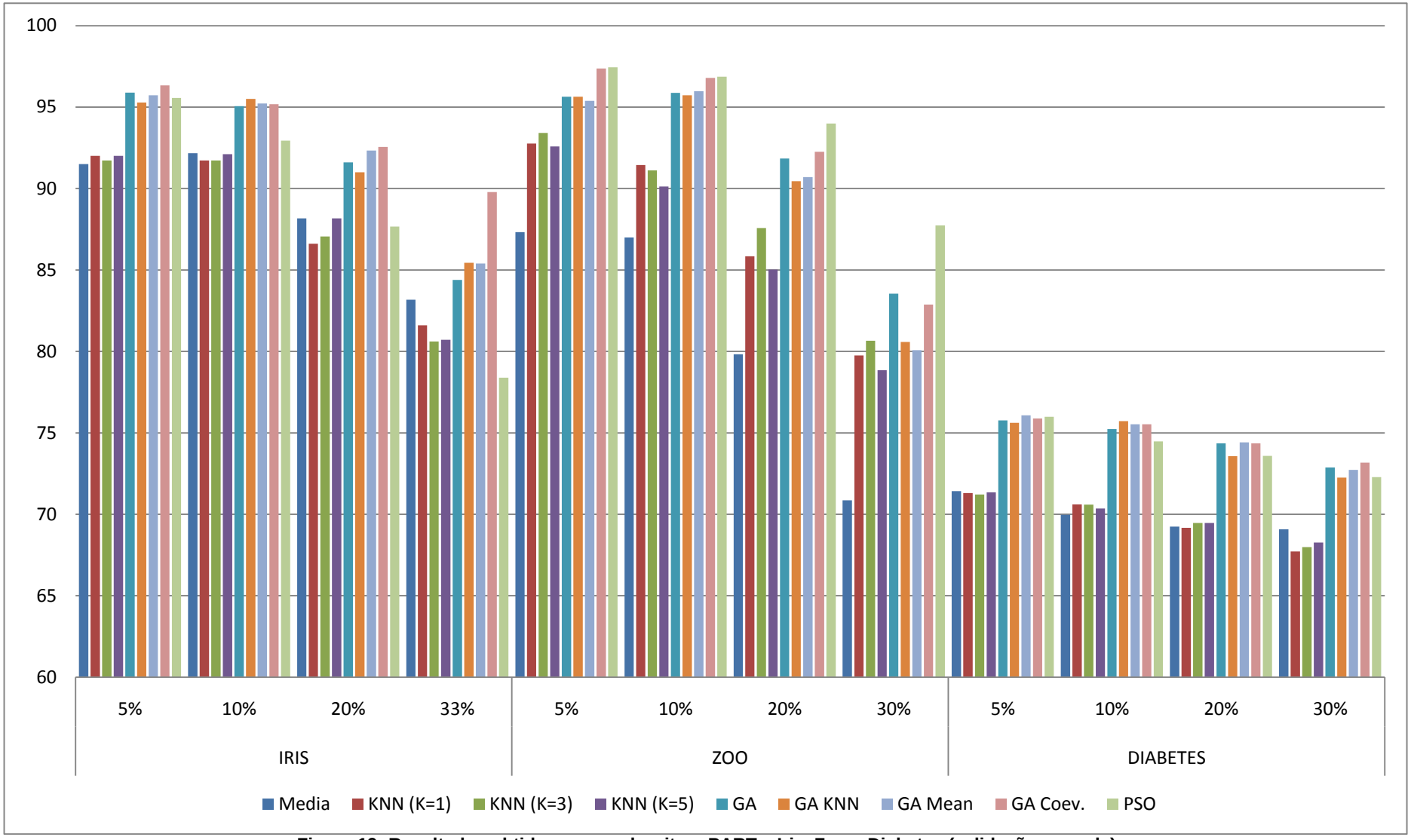

Figura 19: Resultados obtidos com o algoritmo PART – Iris, Zoo e Diabetes (validação cruzada)

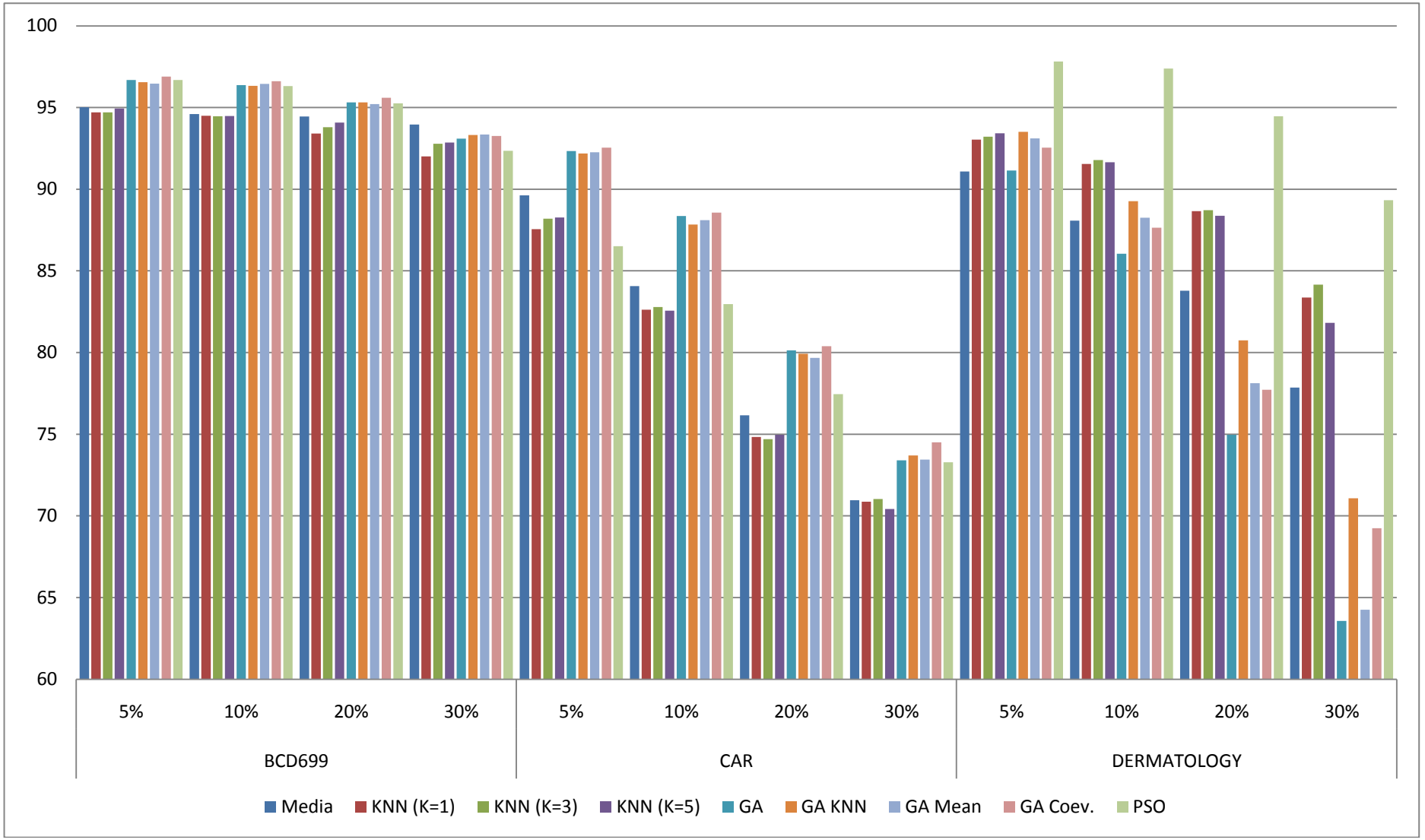

Figura 20: Resultados obtidos com o algoritmo PART – BCD699, Car e Dermatology (validação cruzada)

Tabela 9: Tempo de pré-processamento (em segundos) para o algoritmo PART

	5%	10%	20%	33%	5%	10%	20%	33%
Método	**Tempo de Execução**							
	IRIS				**BCD699**			
Média	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,60±0,69	2,73±0,64	3,60±0,53	5,67±0,58
KNN (K=1)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,67±0,61	3,00±0,00	3,07±0,12	4,00±0,00
KNN (K=3)	1,00±0,00	1,00±0,00	1,07±0,12	1,00±0,00	2,20±0,35	3,00±0,00	3,20±0,20	4,00±0,00
KNN (K=5)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	2,40±0,53	2,93±0,12	3,07±0,12	4,00±0,00
GA	42,93±2,20	43,47±2,05	45,27±2,84	48,53±5,03	74,87±0,31	81,40±2,51	102,73±1,72	147,07±1,15
GA KNN	39,73±1,10	40,20±1,60	42,33±1,62	45,20±1,11	76,60±0,35	82,67±2,00	103,53±1,62	144,07±1,42
GA Mean	42,73±2,23	43,67±2,50	45,40±1,64	48,67±0,61	74,80±0,35	80,73±2,02	102,53±1,63	144,73±0,92
GA Coev.	58,80±5,21	85,93±7,40	111,20±3,86	116,27±2,69	333,40±21,32	439,93±11,37	722,00±20,29	1.449,60±164,45
PSO	111,73±1,67	124,80±6,47	100,13±4,45	78,93±5,83	211,20±17,49	237,33±3,83	322,13±11,29	470,47±16,41
	ZOO				**CAR**			
Média	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	6,53±0,95	7,73±0,92	10,33±1,10	13,13±0,76
KNN (K=1)	1,00±0,00	1,07±0,12	1,00±0,00	1,00±0,00	5,00±0,00	5,20±0,20	6,00±0,00	6,73±0,12
KNN (K=3)	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	5,00±0,00	5,13±0,23	6,07±0,12	6,87±0,12
KNN (K=5)	0,67±0,58	0,67±0,58	0,67±0,58	1,00±0,00	5,00±0,00	5,13±0,23	6,07±0,12	6,93±0,12
GA	44,07±2,77	44,27±2,05	45,93±2,55	47,53±2,89	118,67±2,60	141,27±2,60	198,27±2,37	282,87±2,41
GA KNN	43,07±0,81	43,87±0,81	45,67±1,21	48,67±2,14	123,87±1,75	147,00±3,65	204,13±2,20	286,53±1,10
GA Mean	44,87±0,99	47,73±1,70	51,07±5,51	53,73±4,65	118,13±2,30	139,53±3,41	197,47±2,77	278,60±0,92
GA Coev.	58,33±3,45	95,13±0,50	115,60±2,77	128,80±2,42	1.016,80±25,93	1.186,40±21,81	1.693,47±62,03	2.357,87±126,30
PSO	133,07±3,44	146,67±8,84	160,53±8,27	227,40±12,45	937,87±19,64	1.052,13±22,60	1.262,00±31,59	1.532,00±3,49
	DIABETES				**DERMATOLOGY**			
Média	2,87±0,12	3,00±0,00	4,00±0,00	6,00±0,00	2,13±0,12	3,00±0,00	4,00±0,00	6,00±0,00
KNN (K=1)	3,00±0,00	3,00±0,00	3,40±0,20	4,07±0,12	2,07±0,12	3,00±0,00	3,93±0,12	5,00±0,00
KNN (K=3)	3,00±0,00	3,00±0,00	3,20±0,20	4,00±0,00	2,20±0,20	3,00±0,00	4,00±0,00	5,00±0,00
KNN (K=5)	3,00±0,00	3,00±0,00	3,33±0,31	4,00±0,00	2,20±0,20	3,00±0,00	4,00±0,00	5,00±0,00
GA	104,73±4,40	110,00±1,93	125,07±2,23	151,60±1,64	95,33±0,92	108,53±1,40	159,20±1,83	252,20±2,12
GA KNN	108,27±3,52	113,60±2,55	128,40±2,25	151,40±2,11	99,73±3,59	113,47±2,50	157,67±0,81	241,00±0,53
GA Mean	167,07±105,13	173,93±111,29	185,47±109,58	223,80±133,04	94,27±1,81	107,60±1,22	154,33±0,64	244,20±1,40
GA Coev.	534,27±48,83	639,33±19,00	947,53±32,83	1.874,00±378,70	464,07±14,91	626,80±15,52	938,00±34,73	1.554,87±133,20
PSO	248,53±33,36	307,33±43,35	412,53±48,45	511,20±13,68	612,60±237,32	944,07±311,77	2.535,13±524,16	4.652,73±456,15

4.3. Conclusão

Neste capítulo, foram apresentados e discutidos os experimentos realizados com o intuito de avaliar o desempenho das técnicas bio-inspiradas para fins de imputação de valores faltantes em bases de dados. Os resultados obtidos apontam que, em geral, essas técnicas são as mais eficazes em termos de validação cruzada, porém, menos eficientes em termos computacionais. Os melhores resultados obtidos nas seis diferentes bases e, com os cinco diferentes classificadores, variaram entre as várias técnicas utilizadas. Entretanto, as técnicas utilizando co-evolução e PSO se sobressaíram na maioria dos experimentos. Observou-se também que em algumas circunstâncias as variações de mutação do GA podem trazer benefícios, incorrendo em uma maior eficácia e sem prejudicar sua eficiência.

5. CONSIDERAÇÕES FINAIS

O estudo de uma abordagem de imputação de valores faltantes por otimização numérica utilizando heurísticas bio-inspiradas constitui o propósito e a principal contribuição deste trabalho. Além da adoção de algoritmos genéticos, propõe-se neste contexto a adoção de dois outros algoritmos bio-inspirados que ainda não vêm sendo muito explorados para fins de pré-processamento de dados: a co-evolução cooperativa e a otimização por enxame de partículas (PSO). É feita uma comparação experimental sistemática dessa abordagem de imputação por otimização com duas técnicas tradicionais de imputação adotadas amplamente na literatura (Batista e Monard 2003).

As técnicas bio-inspiradas se mostraram, na grande maioria dos experimentos, bastante eficazes no que diz respeito à etapa de validação cruzada, que dá uma boa noção do impacto da variação dos dados sobre o desempenho dos classificadores. Para todos os seis tipos de classificadores estudados, as técnicas bio-inspiradas se mostraram interessantes para realizar a imputação dos valores faltantes, muitas vezes prevalecendo sobre as convencionais, inclusive na etapa de teste. Contudo, nenhuma das técnicas alcançou a unanimidade dentre os experimentos, não se podendo apontar uma técnica ótima para todos os casos. Se fosse necessário se eleger alguma, poder-se-ia destacar as técnicas de imputação utilizando co-evolução e PSO. Ambas alcançaram muitas vezes resultados expressivos frente às outras técnicas, inclusive à que utiliza GA, notadamente nas bases mais complexas. Contudo, em termos de eficiência, o uso do PSO e da co-evolução apresentou os piores índices de desempenho, com tempos incorridos no pré-processamento bem acima dos apresentados pelas técnicas tradicionais e as técnicas utilizando GA. Isso decorre do fato de as técnicas bio-inspiradas lançarem mão da estratégia *wrapper* importada da linha de seleção de atributos (Kohavi e John 1997) (Freitas 2002).

Com os experimentos realizados, notou-se ainda uma sutil evolução da técnica que utiliza o GA quando se adota os operadores de mutação com KNN ou média/moda. Foram mais eficazes sem perder eficiência. Em alguns casos, o GA convergiu de maneira até mais rápida. Uma linha de investigação futura, no caso do uso do GA com a mutação KNN, seria variar o valor do número de vizinhos *K* ao longo do processo evolutivo, o que poderia trazer ganhos de eficiência e eficácia.

Outra importante linha de pesquisa futura diz respeito à variação na configuração da técnica co-evolutiva. Poder-se-ia variar o tamanho da memória compartilhada ao longo do processo evolutivo, assim como o modo de escolha dos indivíduos para cálculo do valor de *fitness*. Outro ponto passível de ser atacado seria a co-evolução envolvendo algoritmos heurísticos distintos, por exemplo, GA e PSO. Como esses algoritmos seguem estratégias de busca diferentes, sua cooperação poderia trazer benefícios para resolução dos problemas de imputação. Ainda nesta linha de co-evolução, estudos envolvendo processamento distribuído, com utilização de *grids* por exemplo, seriam uma alternativa para melhorar o desempenho desta técnica.

Por fim, vale mencionar outra linha de investigação futura, que já se encontra em curso e envolve o estudo da co-evolução heterogênea para tratar variadas tarefas de pré-processamento de dados. Neste estudo, diversos problemas de pré-processamento de dados seriam tratados de forma simultânea com a co-evolução cooperativa. Acredita-se que, tratando-se de forma automática e cooperativa os diversos problemas de seleção de atributos, seleção de amostras, remoção de ruído e tratamento de valores faltantes, a base de dados final teria um maior nível de qualidade de seus dados, quando comparada com aquela produzida via tratamento seqüencial dos mesmos problemas. Isso porque, em muitas circunstâncias, pode haver uma forte interdependência essas diferentes tarefas de pré-processamento (Pyle 1999).

BIBLIOGRAFIA

Abdella, M., e T. Marwala. “The use of genetic algorithms and neural networks to approximate missing data in database”, *Computing & Informatics* 24(6): 577-589, 2005.

Acuna, E., e C. Rodriguez. “The treatment of missing values and its effect in the classifier accuracy”, D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.), *Classification, Clustering and Data Mining Applications*, 639-648, 2004.

Aha, D., e D. Kibler. “Instance-based learning algorithms”, *Machine Learning* 6(1): 37-66, 1991.

Allison, P.D. “Multiple imputation for missing data: A cautionary tale”, *Sociological Methods & Research* 28: 301-309, 2000.

Asuncion, A., e D.J. Newman. “UCI Machine Learning Repository”, *University of California at Irvine*, 2007. http://ics.uci.edu/~mlearn/MLRepository.html.

Barnard, J., e X.L. Meng. “Applications of multiple imputation in medical studies: From AIDS to NHANES”, *Stat. Methods Med. Res.* 8: 17-36, 1999.

Bashir, S., S. Razzaq, U. Maqbool, S. Tahir, e A.R. Baig. “Using association rules for better treatment of missing values”, *The Computing Research Repository* (CoRR), 2009. http://arxiv.org/abs/0904.3320.

Batista, G.E.A.P.A, e M.C. Monard. “An analisys of four missing data treatment methods for supervised learning”, *App. Artif. Intel.* 17(5):519-533, 2003.

Blum, C., e A. Roli. “Metaheuristics in combinatorial optimization: Overview and conceptual comparison”, *ACM Computing Surveys* 35(3): 268-308, 2003.

Boz, O. “Feature subset selection by using sorted feature relevance”, *Procs. of International Conference on Machine Learning and Applications*, 147-153, 2002.

Brown, M., e J.F. Kros. “The impact of missing data on data mining”, J. Wang (Ed.), *Data Mining: Opportunities and Challenges*, 1 ed., cap. VII, 174-198. Hershey, PA, USA: IGI Publishing, 2003.

Carlisle, A., e G. Dozier. “An Off-The-Shelf PSO”, *Procs. of the Particle Swarm Optimization Workshop*, 1-6, 2001.

Cartwright, M.H., M.J. Shepperd, e Q. Song. “Dealing with missing software project data”, *Procs. of the 9th International Symposium on Software Metrics*, 154-165, 2003.

Clerc, M., e J. Kennedy. “The particle swarm – explosion, stability and convergence in a multidimensional complex space”, *IEEE Transactions on Evolutionary Computation*, 6: 58-73, 2002.

Coelho, A.L.V. “Evolução, simbiose e hibridismo aplicados à engenharia de sistemas inteligentes modulares: Investigações em redes neurais, comitês de Máquinas e sistemas multiagentes”, Tese de Doutorado, FEEC-Unicamp, 2004.

de Castro, L.N. “Fundamentals of natural computing: An overview”, *Physics of Life Reviews* 4: 1-36, 2007.

Eberhart, R.C., e Y. Shi. “Comparison between genetic algorithms and particle swarm optimization”, *Procs. of Evolutionary Programming VII*, 611-616, Springer, 1998.

Eberhart, R.C., P. Simpson, e R. Dobbins. *Computational Intelligence PC Tools.* Morgan Kaufmann, 1996.

Farhangfar, A., L. Kurgan, e J. Dy. “Impact of imputation of missing values on classification error for discrete data”, *Pattern Recognition* 41: 3692-3705, 2008.

Feng, H.A.B., G.C. Chen, C.D. Yin, B.B. Yang, e Y.E. Chen. “A SVM regression based approach to filling in missing values”, *Procs. of Knowledge-Based Intelligent Information and Engineering Systems* (KES05), Lecture Notes in Computer Science 3683: 581-587, 2005.

Frank, E., e H. Ian. “Generating accurate rule sets without global optimization”, *Procs. of Fifteenth International Conference on Machine Learning*, 144-151, 1998.

Freitas, A.A. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* Springer-Verlag, 2002.

Gaul, W., D. Banks, L. House, F.R. McMorris, e P. Arabie. *Classification, Clustering and Data Mining Applications.* Berlin-Heidelberg: Springer-Verlag, 2004.

George, H.J., e P. Langley. "Estimating continuous distributions in Bayesian classifiers", *Procs. of Eleventh Conference on Uncertainty in Artificial Intelligence*, 338-345, 1995.

Grzymala-Busse, J.W. "On the unknown attribute values in learning from examples", *Procs. of the 6th International Symposium on Methodologies for Intelligent Systems*, Lecture Notes in Artificial Intelligence 542: 368-377, 1991.

Grzymala-Busse, J.W., e L.K. Goodwin. "Handling missing attribute values in preterm birth data sets", *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (RSFDGrC 2005), Lecture Notes in Computer Science 3642: 342-351, 2005.

Grzymala-Busse, J.W., e M. Hu. "A comparison of several approaches to missing attribute values in data mining", *Rough Sets and Current Trends in Computing : Second In- ternational Conference* (RSCTC 2000), Lecture Notes in Computer Science 2005: 378-385, 2001.

Han, J., e M. Kamber. *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann, 2001.

Harpham, C., W. Dawson, e R. Brown. "A review of genetic algorithms applied to training radial basis function networks", *Neural Comp. & App.*13(3):193-201, 2004.

Haykin, S. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 1998.

Holland, J. H. *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press, 1975.

Hollander, M., e Wolfe, D.A. *Nonparametric Statistical Methods*. 2a. ed., Wiley, 1999.

Hruschka, E., E. Hruschka, e N. Ebecken. "Missing values imputation for a clustering genetic algorithm", *Procs. of International Conference on Advances in Natural Computation*, Lecture Notes in Computer Science 3612: 245-254, 2005.

Hruschka, E.R., E.R. Hruschka JR., e N.F.F. Ebecken. “A nearest-neighbor method as a data preparation tool for a clustering genetic algorithm”, *Anais do 18 Simpósio Brasileiro de Banco de Dados* (SBBD), 319-327, 2003.

Kennedy, J. “The behavior of particles”, *Procs. of Evolutionary Programming VII* 581-590, Springer, 1998.

Kennedy, J., e R.C. Eberhart. “Particle swarm optimization”, *Procs. of IEEE International Conference on Neural Networks*, 1942-1948, IEEE Press, 1995.

Kennedy, J., R.C. Eberhart, e Y. Shi. *Swarm Intelligence.* Morgan Kaufmann, 2001.

Kennedy, J., e R. Eberhart. “A discrete binary version of the particle swarm algorithm”, *Procs. of IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, 4104-4108, 1997.

Kohavi, R., e G. H. John. “Wrappers for feature subset selection”, *Artificial Intelligence* 97(1-2):273-324, 1997.

Li, D., J. Deogun, e W. Spaulding. “Towards missing data imputation: A study of fuzzy K-means clustering method”, *Rough Sets and Current Trends in Computing*, Lecture Notes in Computer Science 3066: 573-579, 2004.

Libralon, G.L., A.C.P.L.F de Carvalho, e A.C. Lorena. “Pre-processing for noise detection in gene expression classification data”, *Journal of the Brazilian Computer Society* 15(1):3-11, 2009.

Linden, R. *Algoritmos Genéticos.* Rio de Janeiro: Brasport, 2006.

Liping, Z., Y. Huanjun, e H. Shangxu. “A new approach to improve particle swarm optimization”, *Procs. of Genetic and Evolutionary Computation* (GECCO), Lecture Notes in Computer Science 2723:134-139, Springer, 2003.

Little, R.J., e D.B. Rubin. *Statistical Analysis with Missing Data.* New York: John Wiley and Sons, 1987.

Liu, P., e L. Lei. “A review of missing data treatment methods.” *Int. Journal of Intel. Inf. Manag. Syst. and Tech.* 1(3): 412-419, 2005.

Mitchell, T. *Machine Learning*. New York: McGraw-Hill, 1997.

Mundfrom, D.J., e A. Whitcomb. "Imputing missing values: the effect on the accuracy of classification." *Multiple Linear Regression Viewpoints* 25(1): 13-19, 1998.

Potter, M., e K. De Jong. "A cooperative coevolutionary approach to function optimization", *Procs. of Third International Conference on Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 866: 249-257, 1994.

Potter, M., e K. De Jong. "Cooperative coevolution: An architecture for evolving coadapted subcomponents", *Evolutionary Computation* : 8(1):1-29, 2000.

Potter, M., e K. De Jong. "The coevolution of antibodies for concept learning", *Procs. of Fifth International Conference on Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 1498: 530-539, 1998.

Puppala, N., S. Sen, e M. Gordin. "Shared memory based cooperative coevolution", *Procs. of IEEE International Conference on Evolutionary Computation*, 570-574, 1998.

Pyle, D. *Data Preparation for Data Mining.* San Francisco, CA: Morgan Kaufmann Press, 1999.

Quinlan, J.R. *C4.5: Programs for Machine Learning.* San Francisco, CA: Morgan Kaufmann, 1992.

Roberts, M., e E. Claridge. "Cooperative coevolution of image feature construction and object detection", *Procs. of International Conference on Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 3242: 902-911, 2004.

Roth, P. "Missing data: A conceptual review for applied psychologists", *Personnel Psychology* 47: 537-560, 1994.

Rubin, D.B. "An overview of multiple imputation", *Procs. of the Section on Survey Research Methods*, 79-84, 1988.

Schafer, J.L., e J.W. Graham. "Missing data: Our view of the state of the art", *Psychological Methods*7(2):147-177, 2002.

Shi, Y., e R.C. Eberhart. “A modified particle swarm optimizer”, *Procs. of IEEE Conference on Evolutionary Computation*, 69-73, 1998.

Shi, Y, e RC Eberhart. “Parameter selection in particle swarm optimization”, *Procs. of Evolutionary Programming VII* ,611-616, Springer, 1998.

Shi, Y., e R. Eberhart. “Fuzzy adaptive particle swarm optimization”, *Procs. of the 2001 IEEE Congress on Evolutionary Computation*, 101-106, 2001.

Shi, Y., e R. C. Eberhart. “Experimental study of particle swarm optimization”, *Procs. of the World Multiconference on Systemics,Cybernetics and Informatics*, 2000.

Twala, B., M. Cartwright, e M. Sherpperd. “Comparison of various methods for handling incomplete data in software engineering databases”, *Procs. of International Symposium on Empirical Software Engineering*, 105-114, 2005.

Wiegand, R.P., W.C. Liles, e K. De Jong. “An empirical analysis of collaboration methods in cooperative coevolutionary algorithms”, *Procs. of the Genetic and Evolutionary Computation Conference*, 1235-1245, Morgan Kaufmann, 2001.

Witten, I., e E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, 2nd. edition.* San Francisco: Morgan Kaufmann, 2005.

Yang, J., e V. Honavar. “Feature subset selection using a genetic algorithm”, *IEEE Intelligent Systems and their Applications* 13(2):44-49, 1998.

Yoshida, H., K. Kawata, Y. Fukuyama, e Y. Nakanishi. “A particle swarm optimization for reactive power and voltage control considering voltage stability”, *Procs. of Intelligent System Applications to Power Systems*, 177-121, 1999.

Zhang, S., C. Zhang, e Q. Yang. “Guest editors’ introduction – Information enhancement for data mining”, *IEEE Intelligent Systems*, 19(2):12-13, 2004.

Zhang, S., Q. Yang, e C. Zhang. “Data preparation for data mining”, *Applied Artificial Intelligence*, 17(5-6): 375–382, 2003.

Apêndice I – Resultados dos experimentos computacionais

As tabelas abaixo apresentam os resultados obtidos via os experimentos descritos no corpo deste texto. Optou-se por dividir os resultados por algoritmo de classificação utilizado, taxa de atributos faltantes, e técnica de imputação. Em cada tabela, *Crossval* e *Teste* denotam a média obtida de precisão pelo classificador utilizando as partições de validação cruzada e de teste, respectivamente. *Wilcox* traz o *p*-valor do teste de hipótese de Wilcoxon realizado entre o método em questão e a substituição pela média/moda. A coluna *Geração* denota, apenas para as técnicas bio-inspiradas, a geração na qual o melhor indivíduo foi produzido. Em destaque nas tabelas, têm-se os melhores resultados alcançados, em termos de média e desvio-padrão, tanto para a etapa de validação cruzada como para a de teste.

Resultados obtidos com redes neurais RBF para as taxas de 5% e 10% de valores faltantes

	5%					10%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	92,78±1,00		97,11±3,42			93,00±1,09		96,44±3,85		
KNN (K=1)	93,83±1,20	0,26	97,33±2,67	1,00		93,50±0,44	0,56	96,45±3,42	0,84	
KNN (K=3)	93,94±0,92	0,08	97,11±2,69	0,82		93,50±0,50	0,57	96,45±3,42	0,84	
KNN (K=5)	94,17±1,17	0,05	96,89±3,01	0,71		93,28±0,25	0,70	96,22±3,36	0,68	
GA	96,67±0,60	0,00	98,28±1,34	0,17	5,07±1,01	95,11±0,54	0,00	98,71±2,24	0,07	6,47±0,23
GA KNN	96,67±0,58	0,00	98,28±1,97	0,15	6,67±0,50	94,28±0,86	0,11	97,63±1,97	0,34	6,20±1,00
GA Mean	**96,94±0,54**	0,00	98,71±1,71	0,08	6,13±0,70	95,44±1,11	0,00	98,28±1,86	0,13	6,27±0,76
GA Coev.	96,83±0,17	0,00	98,67±1,76	0,15	2,53±1,63	**96,06±0,59**	0,00	**98,89±1,02**	0,11	4,80±0,53
PSO	95,00±0,50	0,00	**98,71±0,65**	0,08	3,87±1,40	92,83±0,17	0,70	98,28±1,62	0,12	4,20±1,40
	ZOO									
Média	93,17±1,64		**95,67±4,04**			91,52±1,98		95,67±4,04		
KNN (K=1)	95,39±1,22	0,01	94,67±3,51	0,49		93,17±1,85	0,15	95,33±2,52	0,80	
KNN (K=3)	94,90±2,40	0,11	95,33±4,51	0,84		92,59±2,43	0,32	93,67±3,06	0,22	
KNN (K=5)	94,57±1,73	0,11	95,00±5,00	0,67		93,00±2,95	0,15	95,00±2,65	0,60	
GA	98,19±0,71	0,00	94,29±2,86	0,92	3,53±1,42	**97,61±0,38**	0,00	95,24±2,86	0,42	4,13±1,63
GA KNN	98,02±0,65	0,00	94,92±4,79	0,75	4,47±1,17	97,53±0,65	0,00	**95,87±2,91**	0,16	4,73±0,64
GA Mean	**98,19±0,62**	0,00	94,92±4,79	0,62	3,13±0,50	**97,61±0,38**	0,00	94,60±3,34	0,57	6,00±0,80
GA Coev.	97,82±0,85	0,00	94,00±4,00	0,26	3,67±0,76	97,03±0,22	0,00	94,67±4,62	0,63	4,47±0,83
PSO	97,28±0,65	0,00	94,60±3,85	0,83	1,27±0,46	97,04±0,49	0,00	94,92±3,85	0,61	3,47±1,21
	DIABETES									
Média	73,39±0,71		77,82±0,54			72,48±0,85		77,25±2,35		
KNN (K=1)	73,03±0,15	0,21	78,87±2,16	0,15		72,24±0,13	0,52	**78,17±1,93**	0,24	
KNN (K=3)	72,60±0,07	0,04	78,47±1,90	0,29		72,25±0,62	0,68	77,91±1,71	0,52	
KNN (K=5)	72,70±0,28	0,04	78,34±2,41	0,32		72,47±0,75	0,98	77,91±1,37	0,50	
GA	75,63±0,92	0,00	78,36±1,04	0,59	8,20±1,06	74,44±0,34	0,00	76,71±1,04	0,53	8,53±1,85
GA KNN	76,04±1,06	0,00	77,36±1,89	0,87	7,87±0,31	74,85±0,94	0,00	76,28±1,05	0,74	6,40±2,36
GA Mean	75,98±0,90	0,00	78,35±1,85	0,62	7,47±0,23	75,00±0,69	0,00	76,54±1,10	0,41	8,13±1,03
GA Coev.	**76,15±0,61**	0,00	**78,87±1,83**	0,23	8,20±0,53	**75,14±0,45**	0,00	76,78±1,53	0,63	8,40±0,53
PSO	75,64±0,27	0,00	78,74±1,39	0,34	6,60±2,25	74,53±0,49	0,00	76,97±0,42	0,74	7,07±0,76

	BCD699									
Média	96,12±0,31		**95,92±1,22**			96,08±0,26		**95,73±0,79**		
KNN (K=1)	96,28±0,40	0,36	95,39±0,63	0,21		95,90±0,32	0,21	95,63±1,34	0,97	
KNN (K=3)	96,20±0,29	0,69	95,63±0,98	0,45		96,21±0,16	0,24	95,49±1,22	0,70	
KNN (K=5)	96,19±0,42	0,72	95,34±0,84	0,14		96,07±0,39	0,95	95,54±1,04	0,67	
GA	97,17±0,41	0,00	95,38±1,15	0,74	7,67±0,42	96,88±0,32	0,00	95,05±1,15	0,56	7,20±1,11
GA KNN	97,02±0,47	0,00	95,14±1,00	0,18	6,67±0,83	96,63±0,34	0,00	95,05±1,79	0,90	6,93±1,90
GA Mean	**97,23±0,41**	0,00	95,33±0,81	0,64	6,20±0,35	96,83±0,13	0,00	94,81±1,45	0,43	7,00±0,20
GA Coev.	97,21±0,50	0,00	95,49±1,40	0,49	6,47±0,81	**96,95±0,42**	0,00	94,92±1,34	0,12	6,87±0,90
PSO	96,69±0,53	0,01	95,19±1,43	0,65	4,00±1,80	96,26±0,30	0,35	95,29±1,49	0,83	3,67±0,81
	CAR									
Média	83,64±0,41		87,38±0,79			80,11±0,31		85,45±0,29		
KNN (K=1)	82,69±0,66	0,08	86,67±0,99	0,25		78,88±1,03	0,01	85,08±0,66	0,44	
KNN (K=3)	83,10±0,53	0,19	86,42±0,72	0,18		79,16±0,54	0,03	85,41±0,49	0,77	
KNN (K=5)	83,17±0,90	0,69	86,77±1,16	0,33		79,43±0,22	0,11	85,94±0,85	0,57	
GA	**87,47±0,44**	0,00	87,51±0,29	0,71	8,80±0,60	84,01±0,39	0,00	85,51±0,58	0,71	7,87±0,76
GA KNN	87,04±0,25	0,00	87,78±0,59	0,32	7,60±0,72	83,65±0,59	0,00	86,61±0,90	0,10	6,40±0,92
GA Mean	87,00±0,26	0,00	**87,82±0,56**	0,30	8,20±0,20	83,78±0,24	0,00	86,28±0,71	0,08	8,47±1,01
GA Coev.	87,25±0,57	0,00	87,61±0,99	0,68	7,20±1,74	**84,14±0,63**	0,00	**87,07±1,84**	0,04	8,40±1,11
PSO	86,24±0,39	0,00	87,11±1,22	0,80	5,73±0,23	83,25±0,45	0,00	85,49±0,91	0,80	5,87±0,12
	DERMATOLOGY									
Média	95,81±0,72		95,43±1,51			94,58±1,40		95,25±0,84		
KNN (K=1)	95,56±0,83	0,77	94,89±1,51	0,72		94,95±0,82	0,68	94,89±2,02	0,67	
KNN (K=3)	95,84±0,58	0,79	95,43±1,98	0,88		95,29±1,24	0,19	94,52±1,19	0,28	
KNN (K=5)	95,72±0,74	0,87	94,79±1,71	0,41		94,74±0,76	0,93	94,15±1,94	0,22	
GA	97,93±0,45	0,00	95,77±1,49	0,26	6,33±2,55	97,47±0,79	0,00	94,68±1,28	0,97	6,00±0,35
GA KNN	98,16±0,30	0,00	95,41±1,64	0,40	6,13±1,94	97,32±0,61	0,00	95,77±1,02	0,22	5,47±0,64
GA Mean	98,09±0,60	0,00	**96,49±1,18**	0,03	5,73±1,55	**97,57±0,57**	0,00	**96,49±1,24**	0,02	5,60±1,22
GA Coev.	**98,20±0,28**	0,00	95,71±0,88	0,73	6,13±0,50	97,50±0,79	0,00	95,71±0,79	0,57	6,13±0,70
PSO	97,84±0,81	0,00	95,41±0,81	0,56	3,00±1,11	97,43±0,68	0,00	95,68±2,15	0,28	5,00±2,12

Resultados obtidos com redes neurais RBF para as taxas de 20% e 33% de valores faltantes

	20%					33%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	88,33±1,33		94,22±4,29			83,95±0,82		91,56±5,00		
KNN (K=1)	89,39±2,30	0,15	96,00±3,71	0,40		84,11±0,48	1,00	92,00±1,76	0,78	
KNN (K=3)	89,22±2,12	0,33	95,78±4,07	0,54		84,05±0,35	0,98	93,11±1,92	0,40	
KNN (K=5)	89,11±2,55	0,50	95,78±3,01	0,62		84,06±0,59	0,85	95,11±1,68	0,04	
GA	90,39±0,54	0,03	98,49±0,37	0,00	5,80±2,43	83,39±0,59	0,57	96,56±0,37	0,00	7,87±0,83
GA KNN	90,56±0,38	0,00	97,63±0,75	0,01	6,67±0,31	83,89±1,46	0,92	**97,42±2,33**	0,00	7,07±1,33
GA Mean	90,50±0,29	0,00	**98,71±0,65**	0,00	6,20±0,53	84,72±0,67	0,60	96,56±2,27	0,00	7,47±1,70
GA Coev.	**92,89±1,39**	0,00	97,78±1,68	0,04	4,73±2,47	**84,83±1,42**	0,80	94,44±1,68	0,14	5,67±3,50
PSO	88,11±0,95	0,57	97,63±1,62	0,02	6,47±0,99	78,94±1,42	0,00	94,84±2,33	0,02	4,93±1,42
	ZOO									
Média	87,57±1,22		94,67±2,89			78,02±1,38		92,67±2,89		
KNN (K=1)	87,90±3,48	0,95	93,67±3,79	0,43		80,00±1,08	0,42	86,33±6,11	0,01	
KNN (K=3)	90,04±0,87	0,17	93,67±3,21	0,67		80,66±2,07	0,38	88,67±4,16	0,04	
KNN (K=5)	89,63±0,85	0,31	94,33±3,06	0,79		80,99±1,54	0,23	89,00±3,61	0,07	
GA	95,23±0,38	0,00	**95,56±2,40**	0,07	7,13±1,36	88,64±2,15	0,00	90,16±3,61	0,97	7,20±1,11
GA KNN	94,90±0,93	0,00	94,60±1,10	0,11	6,07±0,12	87,65±0,74	0,00	93,33±0,95	0,08	5,67±1,03
GA Mean	94,90±1,03	0,00	93,02±0,55	0,85	5,93±1,68	87,33±1,45	0,00	90,48±0,00	0,67	7,93±1,63
GA Coev.	**95,14±0,51**	0,00	93,00±1,73	0,19	5,80±0,20	**89,79±1,79**	0,00	90,67±2,52	0,26	7,53±1,47
PSO	94,49±0,87	0,00	94,60±1,98	0,18	3,27±0,99	86,75±0,51	0,00	**93,97±2,75**	0,02	4,47±1,01
	DIABETES									
Média	71,58±1,32		77,08±1,32			69,89±0,43		75,64±1,27		
KNN (K=1)	70,86±0,73	0,33	77,30±1,45	0,79		69,15±0,10	0,26	76,08±1,71	0,65	
KNN (K=3)	71,45±0,66	0,79	**77,65±1,46**	0,59		69,45±0,46	0,55	**76,38±0,85**	0,42	
KNN (K=5)	71,47±0,75	0,88	77,60±1,74	0,52		69,83±0,93	0,98	76,34±0,91	0,62	
GA	72,47±0,25	0,08	73,64±0,39	0,00	6,93±1,01	70,68±0,42	0,09	71,08±1,51	0,00	7,07±0,81
GA KNN	72,55±0,76	0,07	74,89±2,40	0,20	7,20±0,53	71,12±0,15	0,01	73,33±0,91	0,02	7,60±0,87
GA Mean	72,75±0,90	0,06	73,29±1,65	0,00	7,93±0,83	70,99±0,20	0,03	71,56±0,69	0,00	7,67±0,31
GA Coev.	72,67±0,83	0,08	74,64±2,54	0,08	7,93±0,23	71,54±0,17	0,00	73,86±2,91	0,20	8,40±1,25
PSO	**73,58±0,72**	0,00	76,06±2,07	0,56	7,07±0,12	**72,55±0,98**	0,00	76,32±1,69	0,36	7,73±0,61

	BCD699									
Média	95,87±0,32		95,10±0,72			**94,74±0,52**		94,87±0,65		
KNN (K=1)	94,17±0,20	0,00	94,96±1,14	0,90		92,24±0,26	0,00	94,20±1,08	0,33	
KNN (K=3)	95,48±0,54	0,10	95,39±0,87	0,34		94,08±0,52	0,05	95,25±1,18	0,50	
KNN (K=5)	95,56±0,64	0,34	95,39±1,12	0,19		94,58±0,46	0,56	**95,49±0,71**	0,05	
GA	96,13±0,32	0,10	94,48±1,19	0,56	7,93±1,21	93,55±0,54	0,00	93,81±1,33	0,21	7,13±1,27
GA KNN	95,81±0,38	0,90	**95,43±1,74**	0,11	5,33±1,72	93,42±0,32	0,00	94,81±1,49	0,50	7,47±1,15
GA Mean	95,99±0,39	0,74	95,00±1,65	0,59	6,33±0,76	93,92±0,50	0,00	94,24±2,04	0,77	6,93±1,60
GA Coev.	**96,14±0,19**	0,12	94,44±1,48	0,27	7,13±0,90	93,86±0,40	0,00	94,00±1,90	0,15	8,00±0,40
PSO	94,98±0,24	0,00	94,33±1,57	0,45	4,00±1,44	92,65±0,74	0,00	92,48±1,58	0,00	3,60±1,31
	CAR									
Média	75,13±0,84		82,75±1,12			71,52±0,47		**80,91±2,76**		
KNN (K=1)	74,24±0,25	0,08	81,99±0,23	0,36		71,13±0,60	0,18	77,56±2,12	0,01	
KNN (K=3)	73,82±0,18	0,01	81,78±1,87	0,29		71,51±0,39	0,60	78,82±2,08	0,06	
KNN (K=5)	74,43±0,23	0,25	81,95±1,67	0,26		71,41±0,20	0,47	79,11±3,47	0,22	
GA	77,94±0,24	0,00	82,14±1,33	0,46	8,27±0,64	73,31±0,12	0,00	77,77±0,55	0,00	6,87±1,72
GA KNN	78,21±0,44	0,00	**84,16±1,87**	0,06	7,40±0,87	74,01±0,41	0,00	79,50±1,51	0,21	7,53±1,10
GA Mean	77,47±0,47	0,00	82,95±0,78	0,74	6,73±1,62	73,49±0,18	0,00	78,09±1,65	0,01	8,33±1,01
GA Coev.	**78,29±0,31**	0,00	82,74±0,79	0,72	8,20±0,53	**74,01±0,13**	0,00	79,15±0,69	0,11	8,53±0,81
PSO	77,42±0,44	0,00	82,87±0,41	0,87	6,73±0,81	73,21±0,25	0,00	79,35±0,12	0,18	6,53±0,81
	DERMATOLOGY									
Média	91,63±0,34		94,61±2,46			87,21±0,99		94,34±2,20		
KNN (K=1)	91,15±1,24	0,66	94,25±1,45	0,55		86,71±2,32	0,44	94,79±1,19	0,75	
KNN (K=3)	91,72±0,84	0,87	94,70±1,41	0,97		88,49±0,92	0,01	**94,98±2,06**	0,50	
KNN (K=5)	91,29±0,46	0,71	94,25±1,45	0,65		86,71±1,60	0,30	93,88±1,11	0,67	
GA	93,99±0,83	0,00	94,96±2,17	0,24	6,53±1,90	87,14±1,50	0,74	92,88±1,36	0,34	7,13±1,50
GA KNN	**94,56±0,79**	0,00	**95,77±1,49**	0,07	6,93±0,58	87,71±1,88	0,40	94,50±2,50	0,50	6,87±1,81
GA Mean	94,52±0,56	0,00	94,68±1,90	0,41	6,53±0,50	87,42±1,44	0,63	94,86±1,95	0,30	7,20±1,22
GA Coev.	94,24±0,77	0,00	94,43±2,39	0,90	6,93±0,95	87,71±0,65	0,36	93,88±1,41	0,64	5,67±0,83
PSO	94,33±0,72	0,00	94,96±2,51	0,43	7,13±1,01	**89,03±1,94**	0,04	94,14±3,25	0,83	5,00±2,08

Resultados obtidos com o algoritmo J48 para as taxas de 5% e 10% de valores faltantes

	5%					10%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	92,28±1,55		97,33±1,76			93,11±0,54		97,78±1,92		
KNN (K=1)	92,33±1,20	0,98	97,33±1,76	1,00		92,78±0,69	0,66	97,56±2,14	0,98	
KNN (K=3)	92,00±1,50	0,69	97,33±1,76	1,00		92,72±0,54	0,71	97,56±2,14	0,98	
KNN (K=5)	92,00±1,26	0,70	97,33±1,76	1,00		93,00±0,58	0,82	97,56±2,14	0,98	
GA	95,67±0,29	0,00	97,42±2,58	0,14	4,73±0,31	95,50±0,67	0,01	98,06±1,71	0,53	7,00±0,72
GA KNN	95,61±0,42	0,00	97,85±1,34	0,78	3,33±0,31	94,78±0,48	0,01	98,06±1,71	0,98	5,80±0,87
GA Mean	95,78±0,59	0,00	97,85±1,34	0,00	4,67±0,99	**95,67±0,60**	0,72	98,49±1,62	0,22	6,27±0,46
GA Coev.	**95,89±0,82**	0,00	97,56±1,54	0,03	1,93±1,10	95,50±1,09	0,00	98,00±1,76	0,28	5,20±0,53
PSO	95,22±0,67	0,00	**99,14±0,99**	0,06	4,27±2,86	93,11±0,48	0,09	**98,49±1,34**	0,50	4,13±1,85
	ZOO									
Média	87,49±2,24		92,33±4,73			87,24±0,14		91,00±4,58		
KNN (K=1)	92,26±0,28	0,00	92,00±6,24	0,98		90,78±1,40	0,02	90,67±4,93	0,93	
KNN (K=3)	92,18±1,03	0,00	91,00±7,00	0,72		90,54±1,75	0,02	89,33±4,73	0,45	
KNN (K=5)	92,43±0,87	0,00	91,33±7,23	0,93		91,77±1,17	0,00	90,00±6,24	0,75	
GA	95,64±0,87	0,00	91,43±3,43	0,19	3,80±0,92	95,80±1,73	0,00	88,89±4,89	0,83	7,40±1,40
GA KNN	95,64±1,00	0,00	90,48±5,79	0,98	5,33±1,96	95,80±0,49	0,00	89,52±6,25	0,93	6,60±0,53
GA Mean	95,31±0,65	0,00	91,43±4,36	0,01	2,93±0,81	95,80±0,43	0,00	90,16±5,50	0,00	7,47±1,10
GA Coev.	96,71±0,14	0,00	92,67±4,16	0,40	2,27±0,31	95,88±1,11	0,00	90,33±5,03	0,40	6,73±0,70
PSO	**97,53±0,99**	0,00	**95,24±4,76**	0,64	1,53±0,61	**96,63±1,36**	0,00	**95,87±3,34**	0,42	4,07±0,58
	DIABETES									
Média	72,57±0,75		77,30±1,26			71,44±1,22		76,30±0,53		
KNN (K=1)	72,04±0,75	0,26	77,34±1,98	1,00		71,48±1,21	0,92	76,47±2,16	0,90	
KNN (K=3)	72,58±1,18	1,00	77,17±2,43	0,77		71,42±0,60	0,85	76,08±0,57	0,79	
KNN (K=5)	72,80±0,81	0,62	76,86±2,15	0,65		71,10±2,15	0,60	76,08±0,68	0,85	
GA	77,18±0,49	0,00	76,75±1,15	0,65	7,47±1,62	77,11±0,67	0,00	77,27±2,09	0,30	8,53±0,64
GA KNN	77,00±1,19	0,00	77,23±1,67	0,87	6,67±0,81	76,72±0,29	0,00	77,32±0,86	0,83	7,40±0,72
GA Mean	77,21±0,63	0,00	77,49±1,71	0,05	7,93±1,33	76,56±0,22	0,00	77,14±1,46	0,21	7,07±0,46
GA Coev.	**78,04±0,62**	0,00	77,47±1,64	0,68	8,00±1,40	**77,25±1,22**	0,00	76,12±0,42	0,36	8,27±1,96
PSO	75,44±0,68	0,00	**78,44±0,78**	0,80	5,80±1,51	74,31±0,82	0,00	**77,58±0,49**	0,26	7,07±0,12

	BCD699									
Média	94,19±0,30		93,43±1,08			94,33±0,22		94,10±1,76		
KNN (K=1)	94,19±0,37	0,87	94,10±1,37	0,29		94,25±0,89	0,69	93,43±0,79	0,26	
KNN (K=3)	94,62±0,34	0,13	94,48±1,47	0,11		94,17±0,07	0,65	93,57±1,93	0,53	
KNN (K=5)	94,72±0,09	0,13	93,91±1,58	0,54		94,81±0,24	0,11	93,76±0,95	0,30	
GA	96,45±0,52	0,00	93,52±1,51	0,68	8,47±0,31	96,35±0,64	0,00	93,48±1,01	0,34	7,13±0,95
GA KNN	96,40±0,31	0,00	93,67±2,00	0,59	6,93±0,58	96,12±0,46	0,00	93,33±1,72	0,02	6,07±0,46
GA Mean	96,50±0,27	0,00	92,62±2,00	0,01	6,60±0,20	96,17±0,35	0,00	93,24±1,15	0,21	7,33±1,72
GA Coev.	**96,64±0,44**	0,00	93,09±1,52	0,53	7,93±0,64	**96,65±0,36**	0,00	92,71±1,22	0,26	7,27±0,76
PSO	**96,64±0,46**	0,00	**94,95±1,01**	0,45	3,53±1,85	96,35±0,44	0,00	**94,67±1,57**	0,43	5,93±1,94
	CAR									
Média	87,49±2,24		92,33±4,73			87,24±0,14		91,00±4,58		
KNN (K=1)	92,26±0,28	0,00	92,00±6,24	0,31		90,78±1,40	0,00	90,67±4,93	0,88	
KNN (K=3)	92,18±1,03	0,00	91,00±7,00	0,32		90,54±1,75	0,00	89,33±4,73	0,80	
KNN (K=5)	92,43±0,87	0,00	91,33±7,23	0,59		91,77±1,17	0,00	90,00±6,24	0,41	
GA	95,64±0,87	0,00	91,43±3,43	0,24	3,80±0,92	95,80±1,73	0,00	88,89±4,89	0,12	7,40±1,40
GA KNN	95,64±1,00	0,00	90,48±5,79	0,21	5,33±1,96	95,80±0,49	0,00	89,52±6,25	0,10	6,60±0,53
GA Mean	95,31±0,65	0,01	91,43±4,36	0,00	2,93±0,81	95,80±0,43	0,00	90,16±5,50	0,00	7,47±1,10
GA Coev.	96,71±0,14	0,00	92,67±4,16	0,16	2,27±0,31	95,88±1,11	0,00	90,33±5,03	0,02	6,73±0,70
PSO	**97,53±0,99**	0,00	**95,24±4,76**	0,20	1,53±0,61	**96,63±1,36**	0,00	**95,87±3,34**	0,06	4,07±0,58
	DERMATOLOGY									
Média	89,49±1,85		92,69±4,66			87,48±1,67		92,69±3,57		
KNN (K=1)	91,72±0,60	0,02	92,33±4,11	0,72		90,54±0,44	0,00	93,61±2,33	0,53	
KNN (K=3)	91,49±0,39	0,03	92,78±3,15	0,92		90,97±0,77	0,00	92,69±2,60	0,95	
KNN (K=5)	91,76±0,62	0,01	92,42±3,74	0,75		91,63±1,11	0,00	93,70±3,32	0,31	
GA	92,47±0,28	0,00	95,32±2,18	0,07	8,67±0,81	86,46±1,08	0,23	93,06±1,76	0,41	8,20±0,20
GA KNN	94,65±0,52	0,00	95,50±2,77	0,05	8,73±0,12	90,38±1,59	0,60	**95,50±0,78**	0,90	8,40±0,92
GA Mean	92,86±0,14	0,00	94,32±3,28	0,03	8,33±0,90	87,60±0,68	0,00	93,42±1,80	0,02	8,53±0,42
GA Coev.	93,45±0,60	0,00	**96,16±1,90**	0,21	9,27±0,31	88,42±0,14	0,80	92,69±1,65	0,48	9,07±0,61
PSO	**98,04±0,44**	0,00	95,95±1,43	0,05	4,47±0,81	**97,38±1,09**	0,01	95,14±2,04	0,01	4,60±0,72

Resultados obtidos com o algoritmo J48 para as taxas de 20% e 33% de valores faltantes

	20%					33%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	88,00±1,92		97,11±2,34			83,17±3,71		96,22±1,39		
KNN (K=1)	88,61±1,36	0,79	97,34±2,40	0,87		81,61±1,11	0,43	94,45±2,52	0,37	
KNN (K=3)	88,61±1,34	0,87	96,67±2,40	0,98		81,22±1,02	0,23	95,78±2,14	0,79	
KNN (K=5)	87,72±2,46	0,60	98,00±1,33	0,69		81,00±1,04	0,21	96,67±1,15	0,74	
GA	92,06±1,00	0,00	96,56±0,74	0,57	8,13±0,42	85,83±0,67	0,04	95,48±2,96	0,25	6,47±1,03
GA KNN	92,11±1,17	0,00	97,63±2,27	0,37	5,93±1,62	86,78±2,24	0,00	**96,77±1,71**	0,52	7,00±0,80
GA Mean	**93,11±0,79**	0,95	96,99±2,07	0,16	7,20±1,00	86,33±1,04	0,01	96,13±0,65	0,39	7,00±0,87
GA Coev.	92,33±0,50	0,00	96,44±2,04	0,58	5,87±0,50	**91,33±1,76**	0,05	95,11±1,02	0,09	6,53±1,14
PSO	88,11±0,82	0,00	**98,06±0,65**	0,16	4,53±1,30	78,28±1,07	0,02	96,13±1,94	0,01	3,73±2,14
	ZOO									
Média	79,51±3,04		91,33±1,53			70,45±2,62		88,00±2,00		
KNN (K=1)	85,35±2,10	0,00	90,00±4,00	0,82		79,75±1,78	0,00	88,33±3,06	0,91	
KNN (K=3)	87,16±0,89	0,00	90,00±3,61	0,33		81,15±1,16	0,00	93,00±1,73	0,01	
KNN (K=5)	87,98±0,62	0,00	91,67±1,15	0,85		76,87±3,14	0,01	90,33±3,06	0,20	
GA	91,85±0,74	0,00	90,48±3,30	0,29	8,00±0,53	82,47±1,31	0,00	86,98±1,10	0,87	7,33±0,99
GA KNN	91,03±2,07	0,00	92,38±1,90	0,96	7,40±0,53	81,65±0,62	0,00	85,40±1,98	0,41	6,73±0,70
GA Mean	92,26±1,27	0,00	89,52±3,30	0,00	8,47±0,61	82,30±1,00	0,00	90,16±1,45	0,00	6,53±0,61
GA Coev.	92,84±0,89	0,00	90,67±2,52	0,73	5,60±1,31	84,28±0,79	0,00	86,67±3,51	0,07	6,73±0,61
PSO	**94,07±0,89**	0,00	**94,92±3,34**	0,04	4,73±0,50	**87,32±0,51**	0,00	**92,70±2,20**	0,77	4,80±1,60
	DIABETES									
Média	69,92±0,26		74,99±0,79			68,53±1,83		76,43±2,12		
KNN (K=1)	69,22±1,09	0,29	75,47±2,25	0,62		67,95±1,65	0,48	76,08±0,82	1,00	
KNN (K=3)	69,36±1,14	0,25	75,38±1,06	0,69		68,52±0,49	0,93	76,43±0,53	1,00	
KNN (K=5)	69,50±0,85	0,36	75,42±2,48	0,66		68,78±1,76	0,82	75,51±1,64	0,43	
GA	75,40±0,90	0,00	76,06±1,84	0,28	7,47±0,23	73,37±0,85	0,00	75,45±1,35	0,77	6,20±1,22
GA KNN	74,35±0,85	0,00	76,62±1,12	0,29	6,00±0,69	72,86±0,64	0,00	76,54±1,48	0,40	7,40±0,72
GA Mean	75,20±1,34	0,00	76,41±1,89	0,05	7,40±0,72	**73,85±1,16**	0,00	**76,88±1,40**	0,90	8,93±0,64
GA Coev.	**75,96±1,06**	0,00	76,21±1,93	0,10	8,40±1,11	73,53±0,54	0,00	75,21±1,02	0,48	8,67±0,31
PSO	73,68±0,43	0,00	**76,62±0,13**	0,07	8,07±0,81	72,49±0,52	0,00	76,19±0,72	0,59	8,13±0,46

	BCD699									
Média	94,18±0,68		93,05±1,85			**93,55±0,81**		93,29±1,81		
KNN (K=1)	93,20±0,27	0,02	93,58±1,20	0,39		91,00±0,41	0,00	92,76±0,79	0,29	
KNN (K=3)	93,25±0,69	0,04	93,09±1,66	0,61		91,70±0,37	0,00	92,47±0,22	0,17	
KNN (K=5)	93,12±0,74	0,01	93,14±0,82	0,60		91,61±1,09	0,00	93,09±0,94	0,61	
GA	95,04±0,56	0,02	93,38±1,36	0,23	6,67±1,79	92,55±0,59	0,02	93,19±2,09	0,68	6,47±0,42
GA KNN	94,83±0,48	0,00	93,71±1,43	0,29	6,07±0,42	92,65±0,69	0,66	**93,76±0,50**	0,97	6,40±0,53
GA Mean	95,18±0,27	0,01	93,62±1,37	0,01	6,47±0,61	93,32±1,10	0,00	93,24±1,11	0,28	8,07±1,17
GA Coev.	**95,64±0,29**	0,01	93,53±1,18	0,26	7,87±0,70	93,38±0,18	0,71	93,14±1,76	0,74	7,33±2,47
PSO	95,10±0,08	0,04	**94,62±2,22**	0,10	3,47±1,70	92,35±0,27	0,03	92,29±0,87	0,32	3,40±0,53
	CAR									
Média	79,51±3,04		91,33±1,53			70,45±2,62		88,00±2,00		
KNN (K=1)	85,35±2,10	0,20	90,00±4,00	0,30		79,75±1,78	0,42	88,33±3,06	0,97	
KNN (K=3)	87,16±0,89	0,27	90,00±3,61	0,47		81,15±1,16	0,44	**93,00±1,73**	0,55	
KNN (K=5)	87,98±0,62	0,04	91,67±1,15	0,20		76,87±3,14	0,33	90,33±3,06	0,35	
GA	91,85±0,74	0,00	90,48±3,30	0,80	8,00±0,53	82,47±1,31	0,00	86,98±1,10	0,80	7,33±0,99
GA KNN	91,03±2,07	0,00	92,38±1,90	0,04	7,40±0,53	81,65±0,62	0,00	85,40±1,98	0,90	6,73±0,70
GA Mean	92,26±1,27	0,00	89,52±3,30	0,01	8,47±0,61	82,30±1,00	0,00	90,16±1,45	0,56	6,53±0,61
GA Coev.	92,84±0,89	0,00	90,67±2,52	0,93	5,60±1,31	84,28±0,79	0,00	86,67±3,51	0,62	6,73±0,61
PSO	**94,07±0,89**	0,00	**94,92±3,34**	0,11	4,73±0,50	**87,32±0,51**	0,00	92,70±2,20	0,56	4,80±1,60
	DERMATOLOGY									
Média	84,55±1,17		93,70±2,37			79,98±1,85		91,78±4,53		
KNN (K=1)	88,10±0,28	0,00	92,97±3,12	0,74		84,28±2,45	0,00	92,88±1,19	0,56	
KNN (K=3)	88,83±0,69	0,00	94,25±2,64	0,59		84,14±0,60	0,00	93,52±0,57	0,24	
KNN (K=5)	88,96±0,26	0,00	93,24±3,94	0,95		82,48±1,06	0,01	93,79±1,87	0,21	
GA	78,34±0,85	0,00	89,91±1,80	0,04	8,53±0,23	66,80±0,63	0,00	84,41±2,99	0,00	8,80±0,72
GA KNN	81,05±0,70	0,00	93,24±4,93	0,00	8,20±0,20	69,67±1,10	0,00	90,09±2,73	0,00	8,53±0,12
GA Mean	78,73±0,55	0,00	89,91±1,28	0,34	8,27±0,81	66,44±0,34	0,00	85,50±1,36	0,09	7,53±1,60
GA Coev.	79,27±1,13	0,00	88,86±3,21	0,01	8,80±0,69	68,78±0,85	0,00	85,30±4,91	0,00	7,87±1,10
PSO	**94,70±1,38**	0,00	**94,59±2,74**	0,65	4,67±0,90	**88,94±2,06**	0,00	**93,96±4,13**	0,45	5,47±0,83

Resultados obtidos com o algoritmo IBK para as taxas de 5% e 10% de valores faltantes

	5%					10%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	92,44±1,93		96,67±4,62			91,45±0,67		96,22±3,36		
KNN (K=1)	92,39±1,99	0,92	97,33±4,62	0,54		92,11±0,35	0,35	96,67±3,53	0,68	
KNN (K=3)	92,72±2,02	0,83	97,33±4,62	0,54		91,67±0,29	0,78	96,67±3,53	0,68	
KNN (K=5)	92,78±1,95	0,85	**97,56±4,23**	0,48		91,83±0,58	0,66	96,44±3,91	0,76	
GA	96,78±0,77	0,00	97,20±2,61	0,73	5,20±1,91	96,50±0,73	0,00	95,48±2,58	0,74	7,47±0,70
GA KNN	**97,11±0,77**	0,00	96,99±3,67	0,86	5,47±1,33	**97,00±0,73**	0,00	96,13±2,81	0,46	7,40±0,69
GA Mean	97,00±0,67	0,00	96,99±2,91	0,66	5,93±1,90	96,67±0,44	0,01	95,91±3,78	0,00	7,47±0,70
GA Coev.	96,94±0,82	0,00	97,33±2,67	0,79	4,53±0,46	96,39±1,67	0,00	97,11±2,34	0,55	5,40±2,23
PSO	94,89±0,25	0,00	97,42±1,29	0,68	1,80±1,22	93,39±0,92	0,00	**98,49±0,99**	0,44	5,60±1,44
	ZOO									
Média	95,56±1,62		**95,67±4,04**			93,42±1,00		95,00±3,00		
KNN (K=1)	96,71±0,79	0,07	95,33±3,51	0,80		94,73±0,93	0,07	**96,00±2,65**	0,40	
KNN (K=3)	96,38±0,38	0,28	**95,67±4,04**	1,00		94,74±1,36	0,07	94,67±3,06	0,80	
KNN (K=5)	96,30±0,86	0,31	95,00±5,00	0,67		94,73±1,61	0,07	94,67±2,52	0,79	
GA	97,78±0,89	0,00	95,56±4,50	0,46	2,73±0,81	98,19±0,62	0,00	94,60±1,98	0,15	4,80±0,60
GA KNN	97,70±0,87	0,00	95,24±4,76	0,88	2,07±1,10	**98,44±0,62**	0,00	95,24±1,65	0,38	4,33±0,64
GA Mean	**97,94±0,75**	0,02	95,56±4,79	0,70	2,80±1,00	98,11±0,57	0,00	95,56±3,34	0,05	4,80±0,35
GA Coev.	97,45±0,75	0,00	95,33±5,03	0,46	1,67±0,90	97,20±0,62	0,00	93,67±3,21	0,05	2,27±0,95
PSO	97,28±0,74	0,00	94,60±3,85	0,54	1,07±0,31	96,71±1,17	0,00	95,56±3,34	0,06	2,40±0,72
	DIABETES									
Média	68,78±1,10		72,72±2,68			67,35±0,45		71,24±1,85		
KNN (K=1)	67,59±1,61	0,11	72,72±1,49	0,98		66,43±0,54	0,32	71,55±1,19	0,87	
KNN (K=3)	67,33±1,37	0,08	72,72±1,44	1,00		66,55±0,78	0,35	71,42±1,29	0,97	
KNN (K=5)	67,56±1,39	0,08	72,85±1,45	0,88		66,65±0,78	0,37	71,33±1,06	0,98	
GA	74,19±0,66	0,00	72,12±2,42	0,74	9,20±0,53	73,09±0,12	0,00	70,00±2,54	0,21	9,00±0,53
GA KNN	74,03±0,86	0,00	72,73±1,72	0,34	8,53±0,23	72,78±0,68	0,00	71,04±1,69	0,23	7,93±0,70
GA Mean	74,22±0,36	0,00	72,81±2,86	0,00	9,00±0,72	73,28±0,55	0,00	71,39±3,13	0,00	8,87±0,50
GA Coev.	74,60±0,25	0,00	71,46±2,51	0,90	9,33±0,46	74,60±0,56	0,00	70,24±1,26	0,80	9,40±0,20
PSO	**75,49±0,95**	0,00	**78,48±1,10**	0,80	7,67±0,12	**74,73±0,29**	0,00	**76,88±1,03**	0,90	7,93±0,83

	BCD699									
Média	95,77±0,56		95,06±1,36			95,26±0,81		**95,39±1,23**		
KNN (K=1)	95,14±0,90	0,09	95,01±1,29	0,83		94,54±0,55	0,04	93,91±0,54	0,01	
KNN (K=3)	95,27±0,66	0,10	95,01±1,41	0,92		94,69±0,71	0,15	93,91±0,74	0,01	
KNN (K=5)	95,41±0,81	0,30	95,20±1,36	0,78		94,75±0,53	0,14	93,96±0,66	0,00	
GA	97,52±0,39	0,00	94,86±1,71	0,90	8,33±0,42	97,21±0,74	0,00	94,05±1,94	0,13	8,00±0,72
GA KNN	97,39±0,50	0,00	95,05±1,51	0,69	7,53±1,86	97,18±0,68	0,00	94,38±0,46	0,09	7,47±1,14
GA Mean	**97,63±0,63**	0,01	94,43±0,74	0,36	8,60±0,00	97,45±0,60	0,00	94,00±1,36	0,62	8,00±0,87
GA Coev.	97,62±0,43	0,00	94,82±1,12	0,51	8,20±1,00	**97,69±0,20**	0,00	94,24±1,28	0,12	8,60±0,00
PSO	96,54±0,47	0,00	**95,24±0,95**	0,59	4,60±0,53	96,32±0,41	0,00	94,62±1,75	0,28	4,53±0,81
	CAR									
Média	83,46±0,70		87,54±1,40			78,22±0,97		83,42±1,51		
KNN (K=1)	80,11±0,53	0,00	87,08±0,41	0,59		75,70±0,56	0,00	83,50±0,53	0,77	
KNN (K=3)	80,74±0,61	0,00	86,57±0,73	0,11		76,04±0,17	0,00	83,38±1,34	0,97	
KNN (K=5)	81,09±1,02	0,00	86,52±0,33	0,20		76,56±0,86	0,01	83,02±1,71	0,68	
GA	87,33±0,29	0,00	87,65±1,59	0,56	9,47±0,23	82,94±0,14	0,00	83,28±0,27	0,90	8,93±0,46
GA KNN	**88,60±0,81**	0,00	88,10±0,83	0,43	8,40±1,25	**84,77±0,59**	0,00	84,62±0,64	0,71	8,40±1,20
GA Mean	86,62±0,39	0,00	**88,13±0,85**	0,23	7,93±0,90	82,28±0,53	0,00	83,39±0,68	0,08	8,07±1,30
GA Coev.	87,20±0,18	0,00	87,78±0,48	0,51	9,47±0,12	82,78±0,42	0,00	83,64±0,66	0,00	9,27±0,42
PSO	86,69±0,22	0,00	86,99±1,00	0,18	5,73±1,33	83,30±0,42	0,00	**86,11±1,01**	0,97	6,53±1,21
	DERMATOLOGY									
Média	91,35±0,82		92,78±1,38			88,87±1,92		92,05±1,53		
KNN (K=1)	93,26±1,24	0,00	92,88±1,71	0,61		92,13±1,56	0,00	92,69±0,96	0,39	
KNN (K=3)	92,70±1,25	0,01	92,51±1,94	0,88		92,13±1,11	0,00	92,15±0,57	0,91	
KNN (K=5)	92,74±1,09	0,00	92,78±2,21	0,93		91,42±1,54	0,00	91,41±1,11	0,28	
GA	96,09±0,91	0,00	91,53±0,78	0,48	7,80±1,40	94,49±0,34	0,00	91,26±1,22	0,62	7,87±1,15
GA KNN	95,86±0,77	0,00	91,80±0,87	0,54	7,73±0,90	94,15±0,91	0,00	91,62±1,18	0,82	6,87±0,64
GA Mean	96,06±0,98	0,00	92,43±0,72	0,00	7,80±1,31	94,95±1,07	0,00	91,62±0,71	0,00	7,80±1,11
GA Coev.	96,52±0,87	0,00	93,06±0,42	0,45	8,60±0,35	95,56±0,79	0,00	91,78±1,37	0,93	8,67±0,42
PSO	**97,68±0,80**	0,00	**95,95±1,08**	0,97	4,20±1,11	**97,02±0,52**	0,00	**95,68±1,69**	1,00	4,73±0,61

Resultados obtidos com o algoritmo IBK para as taxas de 20% e 33% de valores faltantes

	20%					33%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	88,28±0,51		94,89±1,92			81,94±1,42		94,22±3,79		
KNN (K=1)	86,50±1,01	0,10	95,56±2,52	0,64		80,72±0,67	0,52	93,56±3,79	1,00	
KNN (K=3)	86,61±0,58	0,06	95,33±2,91	0,75		81,83±0,93	1,00	92,89±3,08	0,45	
KNN (K=5)	86,72±1,29	0,15	95,33±1,76	0,75		80,61±2,10	0,47	94,00±2,40	0,79	
GA	94,50±0,58	0,00	96,34±1,49	0,07	8,67±0,61	86,67±0,87	0,00	93,98±2,69	0,72	8,27±0,64
GA KNN	93,50±0,73	0,00	95,05±2,98	0,50	7,80±1,00	86,67±1,17	0,00	94,19±4,52	0,73	8,00±0,53
GA Mean	93,33±0,60	0,33	96,99±0,37	0,00	8,47±1,14	87,22±1,25	0,14	**96,77±3,35**	0,92	8,33±0,42
GA Coev.	**95,00±0,58**	0,00	96,00±1,76	0,04	6,80±0,53	**90,56±2,47**	0,00	94,00±1,33	0,07	7,33±1,75
PSO	87,22±0,19	0,00	**98,92±0,37**	0,55	5,47±1,36	79,39±0,77	0,01	92,26±5,73	0,29	5,40±0,20
	ZOO									
Média	89,71±1,03		95,67±4,04			79,09±1,24		93,67±2,08		
KNN (K=1)	90,21±1,85	0,63	93,33±3,51	0,13		83,54±1,22	0,00	91,33±1,15	0,08	
KNN (K=3)	91,28±0,71	0,14	94,67±3,21	0,46		83,87±0,75	0,00	93,00±2,65	0,95	
KNN (K=5)	90,62±0,86	0,43	94,67±2,52	0,46		84,45±1,37	0,00	**94,33±3,79**	0,59	
GA	96,54±0,65	0,00	94,60±1,98	0,69	6,13±1,47	89,05±0,14	0,00	90,16±2,75	0,49	8,20±1,11
GA KNN	**97,12±0,71**	0,00	94,60±2,91	0,84	6,33±0,31	87,65±1,08	0,00	92,70±2,40	0,17	6,93±0,61
GA Mean	96,87±0,79	0,00	95,87±2,40	0,83	6,87±0,70	88,72±0,93	0,00	90,79±2,40	0,04	8,07±0,50
GA Coev.	96,95±1,27	0,00	**96,00±3,00**	0,22	6,13±0,42	**91,03±1,79**	0,00	91,00±1,00	0,50	8,87±0,23
PSO	93,99±0,62	0,00	94,60±2,91	0,83	4,20±2,42	87,74±0,29	0,00	94,29±4,15	0,30	3,47±1,55
	DIABETES									
Média	65,27±0,11		70,33±2,20			63,59±0,33		70,46±1,33		
KNN (K=1)	64,08±0,93	0,13	70,28±1,01	0,92		62,17±0,91	0,10	68,80±1,41	0,29	
KNN (K=3)	64,04±0,96	0,07	69,80±2,18	0,77		61,57±1,04	0,02	68,63±1,16	0,12	
KNN (K=5)	63,97±0,37	0,05	68,93±2,47	0,32		60,41±0,33	0,00	69,32±1,13	0,42	
GA	71,38±0,80	0,00	68,53±2,74	0,08	8,00±0,92	68,86±0,39	0,00	66,75±1,91	0,04	7,53±1,10
GA KNN	70,72±0,33	0,00	70,00±1,45	0,30	7,07±0,76	68,38±0,59	0,00	68,05±0,57	0,09	6,47±0,61
GA Mean	71,61±1,10	0,00	68,61±1,55	0,00	8,67±1,10	69,37±0,56	0,00	69,09±0,39	0,00	8,00±1,22
GA Coev.	72,50±0,69	0,00	68,89±2,36	0,30	9,33±0,42	70,53±0,44	0,00	67,93±1,11	0,53	7,40±1,22
PSO	**73,39±1,09**	0,00	**76,28±0,91**	0,97	7,93±1,01	**72,38±0,36**	0,00	**75,19±0,47**	0,07	6,87±0,42

	BCD699									
Média	94,94±0,76		**95,06±1,30**			92,96±0,59		**95,30±0,46**		
KNN (K=1)	92,92±0,30	0,00	94,15±1,93	0,29		84,30±1,82	0,00	94,15±1,04	0,14	
KNN (K=3)	91,68±0,49	0,00	94,20±0,88	0,11		77,87±2,03	0,00	94,58±1,59	0,31	
KNN (K=5)	89,12±0,77	0,00	94,58±0,92	0,25		68,24±2,09	0,00	93,62±1,21	0,01	
GA	96,11±0,41	0,00	93,38±1,66	0,04	8,00±0,40	92,37±0,43	0,12	92,38±1,72	0,00	7,40±0,40
GA KNN	96,32±0,57	0,00	94,81±1,86	0,02	7,27±0,42	**93,27±0,55**	0,44	93,52±1,51	0,00	7,07±0,83
GA Mean	**96,75±0,54**	0,19	94,19±1,01	0,38	8,60±0,69	93,07±1,00	0,04	93,24±0,22	0,00	7,60±1,25
GA Coev.	96,73±0,42	0,00	93,14±2,33	0,21	8,47±1,47	93,12±0,54	0,68	92,52±2,12	0,00	8,53±0,42
PSO	95,26±0,31	0,00	94,33±1,64	0,80	3,73±2,12	92,20±0,50	0,25	91,95±2,07	0,00	3,27±1,29
	CAR									
Média	72,88±0,24		79,71±1,37			69,69±0,62		76,69±1,63		
KNN (K=1)	71,75±0,54	0,00	79,05±1,57	0,51		69,56±0,70	0,93	73,87±0,99	0,00	
KNN (K=3)	71,78±0,32	0,00	78,09±1,14	0,04		69,24±0,81	0,52	72,81±1,49	0,00	
KNN (K=5)	72,47±2,51	0,25	79,94±3,60	0,62		70,42±2,00	0,35	75,15±7,41	0,12	
GA	76,68±0,35	0,00	79,08±0,71	0,62	8,60±0,53	72,10±0,35	0,00	73,08±1,03	0,00	8,00±1,06
GA KNN	77,23±0,28	0,00	77,95±1,22	0,11	9,33±0,12	**73,49±0,80**	0,00	75,18±0,57	0,00	7,20±2,16
GA Mean	76,08±0,31	0,00	77,77±0,93	0,01	7,67±1,10	71,57±0,08	0,00	72,47±0,32	0,09	6,27±0,95
GA Coev.	76,65±0,29	0,00	78,46±0,95	0,00	8,80±0,40	72,21±0,47	0,00	73,02±1,76	0,07	8,73±1,01
PSO	**77,50±0,36**	0,00	**82,08±1,19**	0,00	5,60±0,20	73,13±0,24	0,00	**78,75±0,84**	0,00	6,40±0,92
	DERMATOLOGY									
Média	82,07±1,42		91,69±0,42			72,17±2,50		90,32±2,90		
KNN (K=1)	87,19±1,65	0,00	91,23±0,00	0,58		80,96±3,76	0,00	89,31±0,27	0,56	
KNN (K=3)	87,60±1,20	0,00	91,60±0,57	0,76		80,98±2,47	0,00	89,68±1,24	0,74	
KNN (K=5)	87,94±1,01	0,00	91,96±0,32	0,95		79,68±3,37	0,00	90,04±0,57	0,82	
GA	88,37±0,70	0,00	87,66±0,62	0,00	8,13±1,51	76,47±0,85	0,00	84,82±0,55	0,00	7,60±0,92
GA KNN	88,21±1,15	0,00	88,65±1,40	0,10	6,80±1,06	75,77±1,23	0,00	86,13±1,22	0,00	6,20±2,11
GA Mean	89,03±1,41	0,00	88,47±2,73	0,00	9,07±0,61	76,93±2,22	0,00	86,04±2,03	0,00	7,27±1,30
GA Coev.	89,40±0,73	0,00	90,05±1,11	0,04	8,87±0,23	79,00±1,23	0,00	85,30±1,35	0,04	8,33±0,64
PSO	**94,61±0,94**	0,00	**94,41±1,96**	0,00	6,27±1,01	**89,53±1,69**	0,00	**94,32±1,18**	0,03	5,33±1,01

Resultados obtidos com o algoritmo *Naïve Bayes* para as taxas de 5% e 10% de valores faltantes

	5%					10%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	92,17±0,44		96,67±1,33			91,78±0,75		96,00±1,16		
KNN (K=1)	92,67±0,17	0,51	**98,89±0,77**	0,00		91,33±0,73	0,46	97,34±0,67	0,04	
KNN (K=3)	92,95±0,51	0,17	97,78±1,02	0,10		91,50±0,44	0,56	96,67±0,00	0,25	
KNN (K=5)	92,89±0,54	0,19	98,00±0,67	0,07		91,61±0,67	0,92	96,67±0,67	0,33	
GA	95,83±0,60	0,00	98,49±1,62	0,00	6,13±0,12	94,22±0,69	0,00	98,71±0,65	0,00	7,73±1,03
GA KNN	95,67±0,33	0,00	98,28±0,37	0,01	5,47±0,12	93,78±1,51	0,00	**99,35±0,65**	0,00	5,80±1,40
GA Mean	95,83±0,29	0,00	98,28±0,75	0,00	6,93±0,90	94,56±0,54	0,10	98,49±0,75	0,00	7,13±0,76
GA Coev.	**96,50±1,17**	0,00	98,67±0,00	0,00	4,00±2,31	**95,06±1,25**	0,00	98,89±0,77	0,00	5,13±1,70
PSO	95,00±0,60	0,00	98,06±2,24	0,00	3,47±0,64	92,89±0,95	0,01	98,71±1,12	0,00	2,80±0,35
	ZOO									
Média	93,58±1,73		94,33±3,21			92,35±1,48		95,33±3,51		
KNN (K=1)	94,24±1,00	0,40	**95,33±4,51**	0,51		93,09±1,13	0,47	95,67±2,08	0,82	
KNN (K=3)	93,66±2,00	0,71	**95,33±4,51**	0,51		93,09±1,08	0,31	94,67±2,08	0,62	
KNN (K=5)	93,42±1,36	0,97	94,33±5,13	0,98		93,09±1,73	0,37	94,33±2,52	0,46	
GA	**97,86±1,64**	0,00	95,24±3,81	0,09	4,00±1,40	97,94±0,29	0,00	**95,87±2,91**	0,06	5,53±0,58
GA KNN	97,78±1,54	0,00	94,60±2,91	0,83	3,60±0,00	**98,02±0,49**	0,00	95,24±3,43	0,82	4,53±1,63
GA Mean	97,78±1,48	0,00	94,29±2,86	0,12	3,93±1,62	97,94±0,57	0,00	95,56±3,97	0,23	5,27±1,10
GA Coev.	97,61±0,62	0,00	94,67±4,51	0,24	2,87±0,92	97,12±0,38	0,00	95,67±3,06	0,20	3,33±1,03
PSO	97,53±0,86	0,00	95,24±4,76	0,16	1,07±0,64	96,71±0,75	0,00	95,24±3,43	0,23	1,33±1,40
	DIABETES									
Média	74,02±0,56		79,82±2,13			73,22±0,49		79,43±2,60		
KNN (K=1)	74,21±0,65	0,55	**81,13±1,83**	0,12		73,47±0,52	0,42	81,44±1,02	0,03	
KNN (K=3)	74,28±0,73	0,59	80,87±1,63	0,13		73,59±0,52	0,48	81,35±0,99	0,04	
KNN (K=5)	74,12±0,74	1,00	81,05±1,80	0,12		73,58±0,42	0,43	**81,48±0,87**	0,02	
GA	76,47±0,45	0,00	79,91±1,39	0,74	9,07±0,50	75,89±0,41	0,00	80,26±0,81	0,24	8,20±0,35
GA KNN	76,46±0,67	0,00	80,74±1,51	0,60	7,53±1,17	75,47±0,37	0,00	80,69±1,13	0,75	7,53±0,81
GA Mean	76,22±0,70	0,00	80,78±1,84	0,08	8,60±0,69	75,78±0,72	0,00	79,74±1,48	0,01	7,73±0,99
GA Coev.	**76,62±0,40**	0,00	80,13±1,70	0,11	8,00±1,40	**76,65±0,92**	0,00	79,65±1,10	0,51	8,00±1,20
PSO	75,32±0,78	0,00	78,27±1,25	0,11	7,13±0,76	74,43±0,22	0,00	76,75±1,03	0,12	6,80±0,60

	BCD699									
Média	96,03±0,33		95,15±1,17			95,77±0,48		**95,20±1,39**		
KNN (K=1)	96,14±0,40	0,67	95,20±1,49	0,68		95,77±0,31	0,90	95,01±1,28	0,44	
KNN (K=3)	96,13±0,30	0,70	95,16±1,44	0,84		96,00±0,27	0,20	95,11±1,38	0,71	
KNN (K=5)	96,02±0,36	0,57	**95,25±1,41**	0,60		95,87±0,26	0,44	95,11±1,38	0,71	
GA	96,40±0,30	0,01	95,19±1,43	0,09	7,20±0,92	96,50±0,34	0,00	94,95±1,57	1,00	8,00±0,72
GA KNN	96,04±0,37	0,06	95,24±1,49	0,68	5,53±0,81	95,86±0,22	0,00	95,10±1,73	0,61	4,60±1,64
GA Mean	96,11±0,34	0,00	95,24±1,49	0,77	7,73±1,10	96,43±0,36	0,02	95,10±1,33	0,53	8,13±1,10
GA Coev.	96,32±0,36	0,75	95,20±1,50	0,08	7,67±0,23	**96,73±0,13**	0,00	94,96±1,77	0,47	8,20±0,72
PSO	**96,70±0,32**	0,75	94,95±0,97	0,08	5,07±1,68	96,24±0,29	0,44	95,05±1,30	0,40	4,53±0,76
	CAR									
Média	81,49±0,21		84,41±1,41			79,07±0,18		83,46±1,09		
KNN (K=1)	81,13±0,70	0,21	85,36±1,11	0,08		78,56±0,50	0,28	84,41±1,10	0,03	
KNN (K=3)	81,29±0,60	0,52	85,30±1,21	0,16		78,67±0,40	0,23	84,74±0,92	0,01	
KNN (K=5)	81,42±0,92	0,59	85,37±1,26	0,13		79,24±1,60	0,66	86,22±2,19	0,01	
GA	84,08±0,52	0,00	84,93±0,95	0,24	9,00±0,35	81,84±0,36	0,00	83,70±0,76	0,30	9,27±0,12
GA KNN	84,13±0,49	0,00	85,82±1,01	0,18	8,20±0,53	82,04±0,25	0,00	85,59±0,64	0,22	7,87±0,46
GA Mean	84,01±0,50	0,00	84,82±1,48	0,00	8,60±0,72	82,04±0,45	0,00	83,87±0,40	0,00	8,73±0,81
GA Coev.	86,20±1,52	0,00	85,22±0,75	0,41	8,13±0,81	82,27±0,60	0,00	84,02±0,69	0,15	9,47±0,46
PSO	**86,45±0,43**	0,00	**87,38±0,79**	0,01	5,67±0,83	**83,09±0,23**	0,00	**86,26±0,77**	0,00	6,40±0,40
	DERMATOLOGY									
Média	97,16±0,31		96,35±1,82			96,88±0,48		**96,89±1,94**		
KNN (K=1)	97,16±0,28	0,92	96,07±2,30	0,85		96,52±0,63	0,40	96,16±2,14	0,35	
KNN (K=3)	97,11±0,45	0,69	95,98±2,46	0,68		96,50±0,49	0,17	96,26±2,81	0,65	
KNN (K=5)	96,97±0,08	0,39	95,89±3,11	0,97		96,40±0,48	0,06	96,44±2,17	0,56	
GA	98,98±0,12	0,00	95,50±2,33	0,71	7,40±1,00	98,63±0,31	0,00	96,67±1,84	0,38	7,60±1,04
GA KNN	98,93±0,22	0,00	96,04±2,20	0,56	6,53±1,47	**98,91±0,38**	0,00	96,13±2,05	0,38	7,40±0,20
GA Mean	**99,16±0,20**	0,08	96,04±2,42	0,21	6,00±1,11	98,82±0,17	0,45	95,95±2,36	0,36	8,33±0,50
GA Coev.	99,09±0,28	0,00	**96,71±1,92**	0,56	5,87±2,58	98,70±0,18	0,00	96,26±1,98	0,87	5,87±1,51
PSO	97,66±0,65	0,00	96,58±0,41	0,59	3,13±1,29	97,34±1,01	0,00	95,59±1,09	0,93	4,53±0,99

Resultados obtidos com o algoritmo *Naïve Bayes* para as taxas de 20% e 33% de valores faltantes

	20%					33%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	87,78±0,26		94,00±2,91			83,28±0,53		90,89±4,02		
KNN (K=1)	87,94±0,25	0,92	95,78±2,15	0,10		83,06±0,69	0,88	91,55±4,23	0,73	
KNN (K=3)	87,83±0,29	1,00	95,11±2,70	0,30		83,00±0,29	1,00	92,67±4,16	0,31	
KNN (K=5)	87,61±0,35	0,70	96,00±1,77	0,05		83,06±0,38	0,97	93,78±3,29	0,07	
GA	90,56±1,42	0,00	97,85±1,62	0,00	7,33±1,03	83,06±1,69	1,00	**97,63±2,61**	0,00	8,33±0,99
GA KNN	88,78±0,67	0,00	96,77±1,29	0,01	7,53±1,72	82,56±0,42	0,00	93,55±0,65	0,01	7,33±0,64
GA Mean	90,83±0,44	0,44	97,20±0,37	0,00	8,53±0,64	85,67±0,29	0,00	96,99±0,37	0,00	8,87±0,61
GA Coev.	**92,33±1,92**	0,00	97,33±1,33	0,00	6,80±1,11	**88,72±1,97**	0,02	95,56±1,39	0,00	6,20±0,72
PSO	87,17±0,93	0,07	**98,28±1,34**	0,00	4,53±1,67	79,61±0,51	0,68	95,70±1,62	0,02	6,13±0,50
	ZOO									
Média	89,38±1,31		93,67±2,52			82,55±1,64		91,00±1,00		
KNN (K=1)	90,45±2,00	0,17	93,00±1,73	0,43		83,79±1,24	0,18	93,67±1,53	0,06	
KNN (K=3)	90,70±0,87	0,08	94,00±2,00	0,91		84,12±0,14	0,25	**94,00±3,00**	0,05	
KNN (K=5)	90,45±1,03	0,21	94,33±2,52	0,74		84,61±1,64	0,06	93,67±3,06	0,07	
GA	96,79±0,49	0,00	93,65±1,45	0,20	7,27±1,10	91,28±1,17	0,00	92,70±1,98	0,00	8,60±1,11
GA KNN	96,38±1,11	0,00	94,29±3,30	0,58	6,20±1,06	90,37±0,49	0,00	93,97±2,40	0,23	6,60±1,31
GA Mean	96,87±0,51	0,00	94,60±1,98	0,02	6,27±1,45	91,19±1,17	0,00	92,06±1,45	0,00	7,60±1,25
GA Coev.	**97,20±0,75**	0,00	94,67±4,51	0,04	6,00±1,44	**92,51±1,24**	0,00	92,67±0,58	0,03	7,13±0,61
PSO	94,73±0,38	0,00	**95,24±2,52**	0,13	4,33±0,90	88,56±1,43	0,00	92,70±1,45	0,00	4,93±0,42
	DIABETES									
Média	72,25±1,09		78,91±1,99			70,57±0,92		77,04±2,22		
KNN (K=1)	72,62±0,94	0,48	81,13±0,72	0,00		70,26±0,67	0,56	**80,74±0,74**	0,00	
KNN (K=3)	72,76±1,04	0,52	81,22±0,74	0,00		70,53±0,33	0,97	80,70±0,93	0,00	
KNN (K=5)	72,99±1,07	0,26	**81,39±0,42**	0,00		70,82±0,28	0,42	80,70±0,72	0,00	
GA	74,68±0,63	0,00	78,40±1,62	0,93	7,87±0,81	72,99±0,49	0,00	75,80±2,10	0,59	8,73±0,76
GA KNN	74,37±0,65	0,00	78,96±1,01	0,97	7,33±0,23	72,89±0,40	0,00	78,18±1,13	0,01	7,47±0,42
GA Mean	74,78±0,27	0,00	78,61±0,54	0,15	8,13±0,42	73,11±0,51	0,00	77,49±1,05	0,38	8,80±0,20
GA Coev.	**75,39±0,65**	0,00	78,82±1,48	0,97	8,40±0,72	**73,23±0,20**	0,00	74,51±1,12	0,43	8,33±0,42
PSO	73,83±1,03	0,00	77,14±1,62	0,48	7,73±0,31	72,42±0,97	0,00	75,92±1,17	0,13	7,93±1,15

	BCD699									
Média	95,83±0,41		95,20±1,49			94,44±0,50		**95,35±1,63**		
KNN (K=1)	94,42±0,14	0,00	95,30±1,85	0,85		92,07±0,60	0,00	94,87±2,01	0,46	
KNN (K=3)	95,37±0,29	0,02	95,20±1,49	0,93		93,51±0,66	0,04	95,11±1,88	0,61	
KNN (K=5)	95,35±0,29	0,01	95,15±1,30	0,68		94,44±0,56	0,80	95,25±1,32	0,67	
GA	96,25±0,19	0,04	95,57±2,08	0,26	8,60±0,40	94,17±0,42	0,20	93,76±1,57	0,02	8,87±0,76
GA KNN	95,52±0,41	0,01	95,29±1,96	0,85	7,00±0,72	93,51±0,74	0,72	94,90±2,31	0,03	5,67±1,17
GA Mean	95,92±0,31	0,00	**95,81±2,35**	0,93	8,33±0,42	93,99±0,67	0,00	94,81±2,23	0,00	8,07±0,31
GA Coev.	**96,43±0,43**	0,57	95,30±2,30	0,05	8,73±0,31	**94,45±0,37**	0,05	94,00±1,73	0,62	9,47±0,31
PSO	95,15±0,24	0,11	94,71±1,73	0,36	3,33±0,23	92,50±0,47	0,00	91,95±1,51	0,87	5,73±1,15
	CAR									
Média	75,22±0,37		81,02±1,33			72,45±0,07		79,54±2,16		
KNN (K=1)	74,76±0,47	0,23	83,17±0,69	0,01		72,08±0,48	0,56	82,63±3,22	0,03	
KNN (K=3)	74,37±0,20	0,03	83,60±0,84	0,00		72,34±0,27	0,97	83,21±3,22	0,01	
KNN (K=5)	74,78±0,26	0,23	83,96±0,41	0,00		71,88±1,08	0,28	79,88±5,99	0,71	
GA	77,52±0,47	0,00	80,12±0,74	0,09	8,93±0,31	73,86±0,35	0,00	76,26±0,70	0,04	8,73±0,61
GA KNN	**78,14±0,59**	0,00	**84,39±0,96**	0,21	8,33±1,03	**74,77±0,31**	0,00	**83,24±2,55**	0,03	8,40±0,87
GA Mean	77,46±0,67	0,00	80,79±1,27	0,02	9,07±0,23	73,82±0,33	0,02	76,42±0,90	0,65	8,20±0,40
GA Coev.	77,78±0,34	0,00	80,70±0,55	0,26	8,87±0,61	74,57±0,40	0,00	76,49±1,04	0,02	8,67±0,92
PSO	77,65±0,61	0,00	82,68±0,47	0,00	5,80±0,40	73,30±0,32	0,00	79,21±0,50	0,01	5,60±1,06
	DERMATOLOGY									
Média	94,52±0,81		**96,44±2,51**			91,06±1,54		96,44±1,80		
KNN (K=1)	93,40±1,47	0,07	95,43±1,56	0,19		88,69±1,28	0,00	96,35±1,67	0,73	
KNN (K=3)	94,06±0,79	0,40	95,80±2,07	0,38		89,92±1,24	0,08	96,44±2,51	0,92	
KNN (K=5)	94,33±0,69	0,59	95,43±2,02	0,17		89,87±1,19	0,05	96,53±2,02	0,83	
GA	96,11±0,31	0,00	96,13±2,03	0,68	7,67±0,58	90,72±0,84	0,27	95,68±2,66	0,97	7,80±0,60
GA KNN	96,29±0,69	0,00	95,41±2,21	0,52	8,07±0,31	91,31±1,29	0,15	96,22±2,86	0,14	7,80±0,92
GA Mean	**96,77±0,31**	0,45	96,13±2,50	0,10	7,60±0,00	91,06±0,80	0,01	**96,58±1,63**	0,28	7,13±1,03
GA Coev.	96,36±0,57	0,00	95,71±3,18	0,56	7,60±0,20	**92,06±0,89**	0,87	95,25±2,06	0,34	9,20±0,40
PSO	94,77±0,69	0,00	94,68±2,03	0,56	6,00±0,53	89,31±1,26	0,85	95,05±1,92	0,45	6,20±0,35

Resultados obtidos com o algoritmo PART para as taxas de 5% e 10% de valores faltantes

	5%					10%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	91,50±1,42		96,45±1,68			92,17±0,87		98,22±1,68		
KNN (K=1)	92,00±1,32	0,72	96,67±1,76	0,80		91,72±1,11	0,85	97,78±1,02	0,87	
KNN (K=3)	91,72±1,07	0,85	96,45±1,68	1,00		91,72±0,59	0,71	97,78±1,39	0,65	
KNN (K=5)	92,00±0,93	0,64	96,22±1,68	0,98		92,11±0,48	1,00	98,22±1,68	1,00	
GA	95,89±0,59	0,00	98,06±1,71	0,01	6,27±0,64	95,06±1,36	0,00	98,06±1,29	0,53	7,13±1,47
GA KNN	95,28±0,25	0,00	97,63±0,99	0,50	3,13±0,81	**95,50±0,87**	0,00	97,85±1,49	0,52	6,07±0,90
GA Mean	95,72±0,39	0,00	96,77±1,94	0,00	5,53±1,15	95,22±0,92	0,33	97,42±1,29	0,46	5,00±1,25
GA Coev.	**96,33±0,33**	0,00	95,56±1,68	0,23	2,07±0,76	95,17±0,60	0,00	97,33±1,76	0,96	4,53±1,81
PSO	95,56±0,75	0,00	**98,28±1,34**	0,04	4,73±0,64	92,94±0,67	0,00	**98,28±1,34**	0,79	4,47±0,81
	ZOO									
Média	87,32±1,87		91,33±7,23			87,00±0,71		90,33±6,35		
KNN (K=1)	92,76±0,38	0,00	92,00±7,00	0,78		91,44±1,24	0,00	91,00±5,20	0,79	
KNN (K=3)	93,42±0,62	0,00	91,33±8,14	0,93		91,11±1,54	0,00	90,33±3,79	0,93	
KNN (K=5)	92,59±1,37	0,00	90,33±8,96	1,00		90,12±2,20	0,00	94,00±1,73	0,84	
GA	95,64±0,71	0,00	91,75±7,76	0,17	4,67±1,21	95,88±1,03	0,00	93,65±6,34	0,00	6,13±1,67
GA KNN	95,64±0,87	0,00	91,43±6,60	0,75	5,07±2,73	95,72±1,03	0,00	92,06±4,70	0,93	7,27±0,31
GA Mean	95,39±0,14	0,00	90,79±7,70	0,06	5,13±0,99	95,97±0,62	0,00	92,06±4,70	0,00	5,93±1,42
GA Coev.	97,37±0,57	0,00	91,33±4,62	0,09	1,87±1,15	96,79±0,74	0,00	90,67±4,16	0,01	4,53±0,12
PSO	**97,45±0,87**	0,00	**94,60±3,85**	0,05	1,80±1,25	**96,87±1,43**	0,00	**94,92±1,98**	0,01	2,67±0,12
	DIABETES									
Média	71,42±0,47		77,26±2,28			69,97±0,57		75,47±1,74		
KNN (K=1)	71,31±0,79	0,77	77,04±0,87	0,80		70,61±0,22	0,18	75,47±1,44	0,95	
KNN (K=3)	71,22±1,03	0,65	76,51±1,39	0,66		70,59±1,52	0,43	75,08±2,25	0,79	
KNN (K=5)	71,35±0,57	0,76	76,95±1,94	0,82		70,36±0,49	0,68	73,86±1,73	0,24	
GA	75,76±0,42	0,00	76,15±1,64	0,36	6,00±0,87	75,23±0,47	0,00	76,49±2,47	0,41	6,67±1,94
GA KNN	75,62±0,18	0,00	77,23±1,67	0,50	5,27±0,81	**75,72±0,33**	0,00	76,80±1,56	0,40	5,80±0,40
GA Mean	**76,07±0,10**	0,00	76,23±1,96	0,02	6,60±0,20	75,53±0,69	0,00	77,06±2,31	0,16	6,80±0,87
GA Coev.	75,88±0,99	0,00	76,08±2,85	0,34	6,33±0,81	75,53±0,08	0,00	76,60±3,17	0,18	5,93±1,68
PSO	75,98±0,70	0,00	**79,09±1,28**	0,74	7,00±0,72	74,48±0,29	0,00	**77,27±1,40**	0,28	6,13±0,31

	BCD699									
Média	95,02±0,22		93,62±1,71			94,60±0,15		94,24±1,90		
KNN (K=1)	94,71±0,48	0,00	94,53±1,28	0,22		94,50±0,56	0,00	93,77±0,65	0,35	
KNN (K=3)	94,70±0,39	0,00	94,39±1,00	0,38		94,46±0,40	0,00	**94,67±0,86**	0,80	
KNN (K=5)	94,95±0,40	0,00	94,15±1,04	0,63		94,48±0,58	0,00	93,53±1,37	0,27	
GA	96,69±0,38	0,85	94,24±1,90	0,23	6,13±0,31	96,37±0,28	0,01	94,10±0,84	0,97	5,87±1,22
GA KNN	96,55±0,34	0,04	94,05±1,66	0,82	4,67±1,27	96,33±0,64	0,59	94,48±0,87	0,63	5,67±1,01
GA Mean	96,46±0,47	0,00	94,10±2,14	0,02	5,13±0,76	96,45±0,51	0,00	94,38±1,15	0,34	4,87±0,81
GA Coev.	**96,89±0,30**	0,00	93,86±0,94	0,28	6,67±1,21	**96,61±0,46**	0,74	94,58±1,48	0,74	6,00±1,44
PSO	96,69±0,49	0,00	**95,00±1,13**	0,43	4,27±1,03	96,32±0,38	0,17	94,62±1,15	0,51	3,80±1,11
	CAR									
Média	89,62±0,66		94,28±1,24			84,07±0,79		91,92±1,60		
KNN (K=1)	87,55±0,50	0,00	93,50±1,01	0,20		82,63±0,40	0,00	92,56±0,19	0,43	
KNN (K=3)	88,19±0,42	0,00	93,89±1,04	0,55		82,79±0,26	0,01	92,13±0,49	0,95	
KNN (K=5)	88,27±0,74	0,00	93,43±0,19	0,18		82,56±0,04	0,02	91,67±1,06	0,95	
GA	92,33±0,28	0,00	93,99±1,35	0,97	6,93±1,92	88,36±0,46	0,00	**93,72±2,04**	0,01	6,40±0,20
GA KNN	92,19±0,49	0,00	**94,72±0,84**	0,71	6,33±0,31	87,83±0,28	0,00	92,87±1,09	0,13	5,67±1,17
GA Mean	92,26±0,13	0,00	93,68±1,30	0,00	6,67±2,66	88,10±0,22	0,01	93,03±0,63	0,00	7,93±1,10
GA Coev.	**92,55±0,39**	0,00	94,40±0,93	0,51	7,80±0,72	**88,56±0,19**	0,00	93,04±1,79	0,09	7,40±1,78
PSO	86,51±0,27	0,00	86,92±1,05	0,17	6,40±0,87	82,97±0,35	0,00	85,76±0,26	0,18	6,27±1,21
	DERMATOLOGY									
Média	91,08±1,13		92,88±1,42			88,08±1,60		92,88±0,99		
KNN (K=1)	93,04±0,21	0,38	94,79±1,45	0,11		91,54±0,36	0,79	93,33±1,35	0,44	
KNN (K=3)	93,22±0,49	0,16	95,07±2,07	0,08		91,79±1,24	0,65	93,88±2,77	0,14	
KNN (K=5)	93,42±0,34	0,85	95,43±0,69	0,03		91,65±0,37	0,80	95,16±2,21	0,02	
GA	91,15±0,26	0,00	94,86±1,69	0,02	6,60±1,20	86,05±1,46	0,00	94,14±1,80	0,07	6,80±1,25
GA KNN	93,52±0,18	0,00	**96,22±0,97**	0,01	7,20±0,53	89,26±0,60	0,00	**95,41±1,35**	0,23	7,00±0,00
GA Mean	93,11±0,56	0,00	95,14±2,11	0,00	6,80±0,60	88,26±0,79	0,00	93,69±2,72	0,00	8,27±0,46
GA Coev.	92,54±0,21	0,00	96,07±0,96	0,02	7,13±1,70	87,64±1,91	0,00	93,79±2,02	0,07	7,00±0,72
PSO	**97,82±0,48**	0,00	96,04±0,68	0,00	3,60±1,11	**97,38±0,81**	0,00	95,32±1,63	0,01	4,80±1,06

Resultados obtidos com o algoritmo PART para as taxas de 20% e 33% de valores faltantes

	20%					33%				
Técnica	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**	**Crossval**	**Wilcox**	**Teste**	**Wilcox**	**Geração**
	IRIS									
Média	88,17±1,48		96,67±2,00			83,17±2,62		96,22±1,39		
KNN (K=1)	86,61±1,83	0,27	96,67±3,33	0,98		81,61±1,40	0,43	94,00±2,40	0,04	
KNN (K=3)	87,06±2,58	0,36	97,33±3,05	0,49		80,61±0,35	0,25	95,78±2,14	0,84	
KNN (K=5)	88,17±2,35	1,00	96,67±2,40	1,00		80,72±0,92	0,18	94,00±2,91	0,29	
GA	91,61±0,95	0,00	**98,06±1,29**	0,02	7,87±0,99	84,39±3,46	0,26	**98,06±1,71**	0,00	6,33±1,22
GA KNN	91,00±0,73	0,00	96,77±0,65	0,60	6,27±1,21	85,44±1,99	0,00	96,56±0,37	0,43	5,80±0,69
GA Mean	92,33±0,76	0,49	97,85±0,99	0,44	7,13±1,68	85,39±1,51	0,01	96,99±0,99	0,97	5,60±0,80
GA Coev.	**92,56±1,11**	0,00	96,67±1,15	0,06	5,80±1,56	**89,78±1,97**	0,10	94,44±3,42	0,03	6,40±2,00
PSO	87,67±0,60	0,01	96,34±2,61	0,38	5,07±0,76	78,39±2,14	0,20	95,27±1,34	0,08	4,80±0,20
	ZOO									
Média	79,83±0,94		91,33±3,51			70,86±2,61		87,00±3,61		
KNN (K=1)	85,84±2,13	0,01	89,67±5,13	0,59		79,75±1,24	0,00	89,67±2,31	0,17	
KNN (K=3)	87,57±0,51	0,00	90,67±4,93	0,74		80,66±0,87	0,00	**92,67±3,51**	0,01	
KNN (K=5)	85,02±5,22	0,00	92,00±1,73	0,60		78,85±4,48	0,00	90,67±2,08	0,20	
GA	91,85±4,61	0,00	92,70±5,42	0,08	7,60±0,53	83,54±5,49	0,00	89,52±4,36	0,07	6,40±0,92
GA KNN	90,45±1,40	0,00	91,43±2,86	0,34	6,00±1,40	80,58±1,17	0,00	87,94±2,75	0,22	6,53±1,01
GA Mean	90,70±0,75	0,00	92,38±4,15	0,00	8,13±1,15	80,08±2,29	0,00	89,21±6,34	0,00	5,60±0,40
GA Coev.	92,26±0,51	0,00	90,00±3,61	0,01	6,00±0,92	82,88±1,27	0,00	89,33±0,58	0,09	7,33±1,67
PSO	**93,99±1,00**	0,00	**94,60±2,20**	0,08	3,87±1,63	**87,74±1,92**	0,00	90,79±2,20	0,17	5,60±1,11
	DIABETES									
Média	69,25±0,80		76,43±2,03			69,08±1,54		75,38±1,11		
KNN (K=1)	69,17±0,90	0,88	75,47±0,77	0,27		67,72±0,54	0,04	74,90±1,64	0,49	
KNN (K=3)	69,47±0,48	0,79	74,90±0,47	0,13		67,98±0,72	0,18	75,34±1,93	0,98	
KNN (K=5)	69,46±0,50	0,88	75,95±1,81	0,43		68,26±1,00	0,23	75,86±1,01	0,71	
GA	74,35±0,55	0,00	75,32±2,06	0,28	7,53±1,33	72,88±1,29	0,00	74,72±0,74	0,65	8,80±0,40
GA KNN	73,57±0,15	0,00	76,45±2,10	0,38	5,13±0,23	72,26±0,31	0,00	75,24±0,64	0,85	5,20±0,80
GA Mean	**74,42±0,47**	0,00	76,58±0,65	0,74	6,73±0,81	72,73±0,52	0,00	75,76±1,13	0,12	7,93±1,89
GA Coev.	74,36±1,25	0,00	76,12±1,45	0,68	7,53±1,51	**73,17±1,30**	0,00	75,34±1,18	0,84	7,27±0,81
PSO	73,58±0,87	0,00	**77,14±0,91**	0,84	7,87±0,42	72,29±0,43	0,00	**76,49±1,76**	0,84	7,80±0,53

	BCD699									
Média	94,45±0,27		94,10±1,79			**93,96±0,98**		**94,63±1,82**		
KNN (K=1)	93,41±0,38	0,00	93,91±0,60	0,92		92,01±0,32	0,00	92,81±0,94	0,01	
KNN (K=3)	93,80±0,79	0,00	93,38±0,66	0,45		92,79±0,38	0,00	93,72±0,72	0,11	
KNN (K=5)	94,08±0,56	0,00	94,00±1,41	0,95		92,85±0,52	0,00	94,58±1,01	0,74	
GA	95,31±0,41	0,00	94,38±1,22	0,36	5,40±0,53	93,10±0,98	0,00	93,33±1,94	0,16	6,53±2,95
GA KNN	95,31±0,16	0,00	**94,48±1,36**	0,72	6,07±1,01	93,32±0,72	0,00	93,81±1,03	0,04	6,20±1,00
GA Mean	95,21±0,47	0,00	94,43±1,89	0,23	7,67±0,90	93,35±0,45	0,00	93,76±0,92	0,02	7,00±1,04
GA Coev.	**95,60±0,18**	0,00	94,34±1,58	0,30	6,47±0,64	93,26±0,57	0,00	93,19±0,42	0,21	4,87±0,61
PSO	95,26±0,30	0,00	**94,48±1,33**	0,34	3,73±1,33	92,35±0,24	0,00	92,38±1,49	0,24	2,87±0,46
	CAR									
Média	76,16±0,42		86,86±1,80			70,96±0,15		80,31±3,48		
KNN (K=1)	74,83±0,33	0,00	84,27±1,32	0,02		70,87±0,72	0,98	76,93±1,86	0,01	
KNN (K=3)	74,70±0,46	0,00	85,36±2,14	0,18		71,04±0,49	0,95	76,83±0,85	0,01	
KNN (K=5)	74,97±0,54	0,07	85,20±3,13	0,11		70,42±0,21	0,21	74,32±1,07	0,00	
GA	80,14±0,59	0,00	87,92±0,76	0,09	6,73±1,63	73,41±0,23	0,00	81,52±0,34	0,21	7,20±1,31
GA KNN	79,93±0,71	0,00	88,75±1,44	0,01	6,27±0,31	73,70±0,15	0,00	81,02±0,87	0,01	7,00±0,92
GA Mean	79,68±0,62	0,00	88,94±0,61	0,00	8,00±1,44	73,45±0,05	0,00	81,75±0,98	0,65	7,27±0,42
GA Coev.	**80,39±0,65**	0,00	**89,26±0,67**	0,03	8,53±0,23	**74,50±0,66**	0,00	**83,35±1,64**	0,18	5,87±1,42
PSO	77,45±0,39	0,00	82,54±0,40	0,04	6,40±0,40	73,29±0,41	0,00	79,60±1,05	0,53	6,93±0,90
	DERMATOLOGY									
Média	83,78±1,12		91,78±2,39			77,86±1,47		90,41±2,78		
KNN (K=1)	88,65±1,51	0,01	93,33±3,03	0,15		83,37±2,15	0,00	91,14±1,94	0,63	
KNN (K=3)	88,71±0,72	0,04	93,61±2,77	0,11		84,16±0,90	0,01	90,50±2,23	0,97	
KNN (K=5)	88,37±2,28	0,28	93,52±2,49	0,14		81,82±0,44	0,00	90,32±3,49	1,00	
GA	74,97±0,80	0,00	90,72±2,81	0,45	5,53±0,81	63,57±1,02	0,03	84,68±3,42	0,01	7,67±1,42
GA KNN	80,75±0,34	0,00	92,61±1,56	0,10	8,60±1,11	71,08±0,59	0,06	90,36±2,36	0,00	9,27±0,42
GA Mean	78,13±0,71	0,00	91,62±3,12	0,00	8,27±1,29	64,25±0,99	0,00	84,14±4,14	0,00	6,07±0,95
GA Coev.	77,72±1,09	0,01	89,13±1,56	0,87	7,73±0,76	69,24±3,94	0,07	83,11±3,12	0,00	3,93±1,21
PSO	**94,47±0,41**	0,00	**95,32±2,20**	0,40	5,47±1,67	**89,33±1,47**	0,07	**95,32±2,10**	0,77	4,47±1,89

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo