Aplicação de Algoritmos Evolutivos Multiobjetivo na Seleção de Instâncias

Multi-objective Evolutionary Algorithms in Instance Selection

Fabiana Cristina Bertoni
Universidade Estadual de Feira de Santana
Avenida Transnordestina, s/n - Novo Horizonte
CEP 44036-900 - Feira de Santana – Bahia
55 75 31618086
fcbertoni@gmail.com

Matheus Giovanni Pires
Universidade Estadual de Feira de Santana
Avenida Transnordestina, s/n - Novo Horizonte
CEP 44036-900 - Feira de Santana – Bahia
55 75 31618177
mgpires@ecomp.uefs.br

RESUMO

Sistemas de descoberta de conhecimentos em bases de dados e aprendizagem de máquinas prevêem situações, agrupam e classificam padrões, entre outras tarefas. Apesar desses sistemas se preocuparem em gerar informações de fácil interpretação, confiáveis e de forma rápida, as extensas bases de dados normalmente utilizadas dificultam o alcance de precisão aliada a um baixo custo computacional. Para resolver esse problema, as bases de dados podem ser reduzidas com o objetivo de diminuir o tempo de processamento e guardar apenas informações suficientes e relevantes para a extração do conhecimento. Nesse contexto, métodos para reduzir e filtrar as bases de dados vem sendo propostos, com destaque para a Seleção de Instâncias, que seleciona um subconjunto de exemplos que possa ser usado para gerar modelos de classificação com a mesma precisão que os modelos gerados a partir do conjunto original. Nas últimas décadas, diversas abordagens para este fim vêm sendo apresentadas, e dentre elas as que utilizam Algoritmos Evolutivos. Entretanto, apesar dos métodos de seleção de instâncias buscarem otimizar dois objetivos considerados conflitantes entre si, precisão na classificação e redução do custo computacional, apenas um algoritmo para otimização multiobjetivo foi aplicado até o momento neste problema. Assim, este trabalho buscou avaliar o desempenho de Algoritmos Evolutivos Multiobjetivo amplamente conhecidos pela comunidade científica, tais como o NSGA-II e o SPEA-II, na seleção de instâncias. Os resultados, comparados com os disponíveis na literatura correlata, demonstram que os algoritmos NSGA-II e SPEA-II podem ser aplicados no processo de seleção de instâncias para problemas de classificação, apresentando elevadas taxas de redução do número de instâncias e tempo de execução reduzido, sem alterações significativas na precisão.

Palavras-Chave

Seleção de Instâncias; Problemas Multiobjetivo; Algoritmos Evolutivos Multiobjetivo.

ABSTRACT

Systems for Knowledge Discovery in Databases and Machine Learning predict situations, group and recognize patterns, among other tasks. Although these applications are concerned in generate fast, reliable and easy to interpret information, extensive databases used for such applications make difficult achieving accuracy with a low computational cost. To solve this problem, the databases can be reduced aiming to decrease the processing time and facilitating its storage, as well as, to save only sufficient and relevant information for the knowledge extraction. In this context, methods to reduce and filter databases have been proposed, especially the Instance Selection Methods that selects a subset of examples from the original training data. The subset should maintain all the information of the original set, so that it can be used to generate classification models with the same accuracy as models generated by using the original set. In the last decades, several approaches have been suggested and studied in order to select instances, among them the Evolutionary Algorithms. Although the instance selection methods aim to optimize two goals conflicting with each other, accuracy and reduce computational cost, only an algorithm for multi-objective optimization has been applied to this problem. Therefore, the aim of this study is to perform instance selection based on widely known Multi-objective Evolutionary Algorithms, such as NSGA-II and SPEA-II, and to evaluate the classification performance over different domains datasets. The results, compared to others available in the correlative literature, demonstrate that NSGA-II and SPEA-II algorithms can be applied in instance selection for classification problems, because many superfluous instances are removed from the training set, reducing runtime in the classification stages, without significant changes in accuracy.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

CCS Concepts

Computing methodologies → Machine learning → Machine learning approaches → Bio-inspired approaches.

Keywords

Instance Selection; Multi-objective Problems; Multi-objective Evolutionary Algorithms.

1. INTRODUÇÃO

Classificação de padrões consiste em definir uma classe, dentre um conjunto de classes existentes, para cada instância de um conjunto de dados. Este conjunto de dados representa um determinado domínio, em que cada instância é composta por atributos ou características que descrevem os dados. A grande extensão das bases de dados, presente em grande parte dos problemas reais, afeta diretamente a precisão e o custo

computacional, em termos de tempo, de métodos de classificação de padrões. Neste contexto, técnicas para reduzir as bases de dados vêm sendo propostas. A redução de dados pode ser alcançada de diversas maneiras, dentre elas, a Seleção de Características e a Seleção de Instâncias. Ao selecionar características, reduzimos o número de colunas em um conjunto de dados [15] e selecionando instâncias, reduzimos o número de linhas [22]. A Seleção de Instâncias ou Protótipos vem sendo destacada como um dos mais promissores mecanismos de redução de dados [16]; [14].

Mais detalhadamente, selecionar instâncias consiste em obter um representativo subconjunto de dados com tamanho menor e com um desempenho para classificação similar ou até mesmo maior que o original, dada a eliminação de instâncias supérfluas e ruidosas. Além de permitir que os algoritmos de classificação de padrões se concentrem na parte relevante dos dados, melhorando seu desempenho, a seleção de instâncias pode tornar possível a aplicação de tais algoritmos, dada a limitação de capacidade de alguns deles em lidar com bases de dados de tamanho elevado.

Embora sejam muitos os benefícios advindos da seleção de instâncias, é necessário considerar ganhos e perdas, geralmente associados à quantidade de redução que se deseja ter e à qualidade na precisão pretendida [19]. Assim, na seleção das instâncias, existe um problema de *trade-off* entre redução e precisão, onde se tenta manter a qualidade na classificação, minimizando a quantidade de dados. Na busca do melhor equilíbrio entre redução dos dados e precisão, diversas metodologias vêm sendo propostas para selecionar instâncias. No entanto, após o estudo dos trabalhos relacionados, notou-se que quase todos utilizaram algoritmos evolutivos para seleção de instâncias, mas apenas um deles é um Algoritmo Evolutivo Multiobjetivo (AEMO), apesar de tais métodos buscarem otimizar dois objetivos considerados conflitantes entre si, precisão e redução do custo computacional. O único AEMO utilizado foi proposto no trabalho de [7]. No entanto, este algoritmo possui uma característica que o difere dos algoritmos evolutivos multiobjetivo tradicionais, que é calcular um valor de *fitness* tanto para as soluções não dominadas quanto para as dominadas, o que o torna mais dispendioso computacionalmente.

Em suma, de acordo com o levantamento bibliográfico realizado, não foi encontrado até o momento na literatura nenhum trabalho que usasse para selecionar instâncias um dos AEMO amplamente conhecidos pela comunidade científica, tais como o *Non-dominated Sorting Genetic Algorithm* (NSGA-II) [11] e o *Strength Pareto Evolutionary Algorithm* (SPEA-II) [27]. Outro aspecto importante é que a maioria deles utiliza os classificadores K-NN (*K-Nearest Neighbor*) e SVM (*Support Vector Machine*) para avaliar a taxa de classificação. Somente os trabalhos de [13] e [3] utilizaram um Sistema Baseado em Regras Fuzzy, que foi construído a partir do conjunto reduzido de instâncias.

Neste contexto, este trabalho se propõe a avaliar a aplicação dos algoritmos NSGA-II e SPEA-II na seleção de instâncias em bases de dados de classificação, utilizando como classificador o algoritmo K-NN. Os resultados serão comparados com os obtidos sem realizar a seleção de instâncias (ou seja, com o conjunto de dados completo) e com outros resultados apresentados nos trabalhos correlatos disponíveis na literatura.

Este trabalho está organizado como segue. Na Seção 2 é apresentada uma breve descrição de alguns dos trabalhos relacionados encontrados durante o levantamento bibliográfico. A Seção 3 aborda conceitos sobre os Algoritmos Evolutivos Multiobjetivo. A Seção 4 trata da aplicação de Abordagens Evolutivas Multiobjetivo na seleção de instâncias. Na sequência, a Seção 5 descreve os experimentos e apresenta os resultados. Por fim, na Seção 6, são apresentadas as conclusões.

2. TRABALHOS CORRELATOS

Nesta seção são descritos os trabalhos que realizam a seleção de instâncias, encontrados na fase de levantamento bibliográfico desta pesquisa. São apresentadas as metodologias aplicadas e um resumo dos resultados obtidos e das conclusões.

Os autores de [6] compararam o desempenho de quatro algoritmos evolutivos: *Generational Genetic Algorithm* (GGA) [17], *Steady-State Genetic Algorithm* (SSGA) [23], *CHC Adaptive Search Algorithm* (CHC) [12] e *Population-Based Incremental Learning* (PBIL) [4], com algoritmos não evolutivos. Os tamanhos das bases de dados variaram de 148 a 768 instâncias para as bases consideradas pequenas, e de 7200 a 10992 instâncias para as bases consideradas médias. No total foram analisadas 13 bases, sendo 10 pequenas e 3 médias. De acordo com as conclusões, o algoritmo CHC foi o que obteve os melhores resultados. No entanto, os autores fazem uma ressalva em relação ao longo tempo de processamento do CHC, o que se torna um empecilho em aplicá-lo em grandes bases de dados.

No trabalho de [7] é proposto um algoritmo chamado *Intelligent Multi- Objective Evolutionary Algorithm* (IMOEA), o qual é uma versão multiobjetivo do *Intelligent Genetic Algorithm* (IGA) proposto por [18]. O objetivo do artigo é aplicar o IMOEA para simultaneamente selecionar instâncias e características para projetar ótimos classificadores 1-NN, ou seja, obter classificadores que maximizem a taxa de classificação enquanto minimizam a quantidade de instâncias e características. De acordo com os autores, o IMOEA obteve alto desempenho em comparação ao IGA e ao *Strength Pareto Evolutionary Algorithm* (SPEA) [26]. Foram utilizadas 11 bases de dados, as quais possuíam de 150 a 1473 instâncias, e de 3 a 60 características.

Em [3], realiza-se a seleção de instâncias por meio de uma abordagem co-evolutiva, ou seja, um algoritmo genético mono-objetivo é responsável por obter um conjunto de instâncias reduzido, o qual é usado por um AEMO para a construção da base de conhecimento. Os autores mostraram que com um conjunto entre 10% a 20% do conjunto original é possível obter resultados satisfatórios, e além disso, reduzir cerca de 86% do tempo computacional. Os tamanhos das bases de dados variaram de 4052 a 40768 instâncias.

O trabalho de [13] fez um estudo comparativo de 36 técnicas de seleção de instâncias aplicadas em 37 bases de dados de diferentes tamanhos. Dentre as técnicas de seleção há algoritmos evolutivos e não evolutivos. Das 37 bases, 20 foram consideradas pequenas, que variavam de 150 a 1473 instâncias, e 17 foram consideradas médias ou grandes, contendo de 2201 a 19020 instâncias. De acordo com as conclusões, o algoritmo *Relative Neighborhood Graph Editing* (RNG) é indicado para bases pequenas, enquanto que os algoritmos *Intelligent Genetic Algorithm* (IGA) e *Populational Based Incremental Learning* (PBIL) são indicados para as bases médias ou grandes. No entanto, o PBIL apresenta, na média, menores tempos de execução.

Em [21], os autores propuseram um algoritmo genético eficiente, chamado EGA, para selecionar instâncias, o qual utiliza conceitos

da evolução biológica para melhorar seu desempenho, principalmente em relação à diversidade da população. Este algoritmo foi comparado com um AG tradicional e com os seguintes algoritmos: *Iterative Case Filtering* (ICF) [5], DROP3 [24] e IB3 [1]. O EGA superou os quatro algoritmos em termos de média de acurácia, e, além disso, conseguiu melhores taxas de redução com menor tempo computacional. Os experimentos foram realizados em quatro bases de dados, as quais tinham de 149 a 11742 instâncias, e de 200 a 18236 características.

3. ALGORITMOS EVOLUTIVOS MULTIOBJETIVO

Um Problema de Otimização Multiobjetivo (MOOP, do inglês *Multiobjective Optimization Problem*) possui um conjunto de funções objetivo a serem otimizadas (maximizadas ou minimizadas). Além disso, possui restrições que devem ser satisfeitas para que uma solução seja factível para o problema.

As funções objetivo empregadas nos MOOP são em geral conflitantes entre si. Em um MOOP, emprega-se o conceito de Dominância de Pareto para comparar duas soluções factíveis do problema. Dadas duas soluções X e Y, diz-se que X domina Y (denotado como $X \prec Y$) se as seguintes condições são satisfeitas: (a) a solução X é pelo menos igual a Y em todas as funções objetivo; e (b) a solução X é superior a Y em pelo menos uma função objetivo. Assim, existe um conjunto de soluções que possuem vantagens em desempenho, mas que não são melhores em custo e vice-versa. Em um MOOP, o conjunto de soluções não-dominadas é chamado de conjunto Pareto-ótimo, que representa as soluções ótimas do problema. O conjunto de valores das funções objetivo das soluções do conjunto Pareto-ótimo é denominado Fronteira de Pareto. [10] aponta três importantes metas em otimização multiobjetivo. A primeira é encontrar um conjunto de soluções que esteja o mais próximo possível da Fronteira de Pareto. Soluções muito distantes desta fronteira não são desejáveis. A segunda é encontrar a maior diversidade dentro das soluções, assegurando a maior cobertura possível da fronteira. Por fim, dado que encontrar um conjunto de soluções uniformemente distribuído é uma tarefa que pode consumir consideráveis recursos computacionais, é necessário que tais soluções sejam obtidas eficientemente.

Existem técnicas tradicionais para solução de MOOP, como os métodos de Somatório dos Pesos e Programação por Metas [10]. A principal vantagem dessas técnicas é que possuem provas de convergência que garantem encontrar pelo menos uma solução Pareto-ótima [9]; [10]. Entretanto, todas as técnicas tradicionais reduzem um MOOP para um problema de objetivo simples. Cada técnica utiliza uma forma diferente de redução e introduz parâmetros adicionais, sendo que a escolha desses parâmetros afeta diretamente os resultados obtidos. Cada vez que os parâmetros são modificados, é necessário resolver um novo problema de otimização simples. Além disso, alguns métodos não garantem soluções ao longo de toda a Fronteira de Pareto (diversidade das soluções). Como alternativa para superar essas desvantagens, outras técnicas foram propostas para tratar MOOP, e dentre estas, destacam-se os Algoritmos Evolutivos.

Os Algoritmos Evolutivos Multiobjetivo (AEMO) são métodos de busca que imitam os mecanismos de evolução natural das espécies, compreendendo processos de evolução genética de populações, sobrevivência e adaptação dos indivíduos [10]. A aplicação dos AEMO para MOOP apresenta vantagens em relação às técnicas tradicionais [8], como por exemplo, não introduzem parâmetros adicionais para a resolução do problema, trabalham diretamente com várias funções usando o conceito de dominância de Pareto e um conjunto diversificado de soluções pode ser encontrado apenas em uma execução do AEMO. Dentre os AEMO mais conhecidos destacam-se o *Non-dominated Sorting Genetic Algorithm* (NSGA-II) [11] e o *Strength Pareto Evolutionary Algorithm* (SPEA-II) [27].

O algoritmo NSGA-II é baseado em uma ordenação elitista por dominância. Para cada solução i, contida na população de soluções, são calculados dois valores: (a) nd_i, o número de soluções que dominam a solução i; e (b) U_i, o conjunto de soluções que são dominadas pela solução i. As soluções com $nd_i = 0$ estão contidas na fronteira F_1. Em seguida, para cada solução j em U_i, decrementa-se o nd_j para cada $i \prec j$, onde $i \in F_1$. Se $nd_j = 0$, então a solução j pertence à próxima fronteira, neste caso, F_2. Tal procedimento é repetido até que todas as soluções estejam classificadas em uma fronteira. Esse procedimento consiste em classificar as soluções de um conjunto M em diversas fronteiras $F_1,F_2,...F_k$ conforme o grau de dominância de tais soluções. Para garantir a diversidade na fronteira calculada, o NSGA-II emprega uma estimativa da densidade das soluções que rodeiam cada indivíduo da população. Assim, calcula-se a média da distância das duas soluções adjacentes a cada indivíduo para todos os objetivos, denominada de distância de multidão ou *crowdist*. A aptidão de cada solução (indivíduo) i é determinada pelos seguintes valores: (a) $rank_i = k$, o valor de $rank_i$ é igual ao número da fronteira F_k à qual pertence; e (b) $crowdist_i$, o valor de distância de multidão de i. Assim, no processo de Ordenação por Dominância, uma solução i é mais apta que uma solução j se: (a) i possui um ranking menor que j, ou seja, $rank_i < rank_j$; e (b) se ambas as soluções possuem o mesmo ranking e i possui um maior valor de distância de multidão.

No SPEA-II, a aptidão para cada solução i é obtida em várias etapas. Primeiro, calcula-se um valor de aptidão (denotado por $strenght_i$), que representa o número de soluções que são dominadas pela solução i. Calcula-se também o valor de raw_i, que representa o somatório dos valores $strenght_j$ das soluções j que dominam i. Conforme [27], este mecanismo em certo nível ordena as soluções por dominância, mas não enfatiza a preferência de uma solução sobre outra. Para resolver esse problema, o SPEA-II usa uma informação de densidade, chamada de $dens_i$, baseada no método de k-vizinhos, onde a densidade em qualquer ponto é uma função decrescente em relação a k-ésima solução mais próxima. Finalmente, a aptidão final para cada solução i, denotada por A_i, é dada pela soma de raw_i com $dens_i$.

4. ABORDAGENS EVOLUTIVAS MULTIOBJETIVO APLICADAS NA SELEÇÃO DE INSTÂNCIAS

Nesta seção serão detalhados os parâmetros dos algoritmos NSGA-II e SPEA-II utilizados para selecionar instâncias: a codificação dos cromossomos, os operadores genéticos (cruzamento e mutação) e a função de avaliação dos cromossomos (ou cálculo do *fitness*).

A codificação adotada para os cromossomos é uma sequência de zeros e uns, onde o valor 'um' representa a presença de uma determinada instância e o valor 'zero' representa a ausência. Esta codificação também é utilizada pela maioria dos trabalhos apresentados na Seção 2. Como consequência desta forma de

codificação, o tamanho do cromossomo será igual ao tamanho do conjunto de instâncias. O operador de cruzamento escolhido foi o *Half Uniform Crossover* (HUX) [12], devido aos bons resultados obtidos pelo algoritmo CHC, que utiliza este operador, conforme apresentado no trabalho de [6]. Em relação à mutação foi escolhido um operador que simplesmente troca o valor de um gene por outro. Se o valor de um gene é um, então este valor é convertido para zero, e vice-versa. A seleção dos cromossomos é feita pelo operador Torneio [20]. A avaliação dos cromossomos foi feita a partir de dois objetivos, taxa de acurácia e taxa de redução do conjunto de treinamento original, os quais são definidos pelas equações 1 e 2, respectivamente:

$$taxa\_acuracia = \frac{qtdade\_acertos}{total\_exemplos} \quad (1)$$

$$taxa\_reducao = \frac{total\_exemplos - qtdade\_selecionada}{total\_exemplos} \quad (2)$$

A variável qtdade_acertos contabiliza a quantidade de exemplos que foram classificados corretamente, a variável qtdade_selecionada indica a quantidade de exemplos (ou instâncias) selecionadas, e a variável total_exemplos é a quantidade total de exemplos do conjunto de treinamento. O valor da variável qtdade_acertos é resultado da aplicação do classificador K-NN, com K=1. A facilidade de implementação e a efetividade, fazem do K-NN um dos algoritmos mais utilizados em experimentos de classificação [25] e mais especificamente, em problemas de seleção de instâncias, como pode ser observado na Seção 2.

Ao término da execução dos AEMO, a solução final que representa as instâncias selecionadas é a mais próxima do ponto médio da fronteira de Pareto. Este critério foi adotado visando o balanceamento entre os objetivos taxa de acurácia e taxa de redução.

5. EXPERIMENTOS E RESULTADOS

Nesta seção são apresentadas as características das bases de dados usadas nos experimentos, a descrição dos experimentos realizados e a discussão dos resultados obtidos.

5.1 Características das Bases de Dados

Foram consideradas para os experimentos um total de 36 bases, as quais foram divididas em dois grupos, "pequenas" e "médias-grandes", de acordo com o número de instâncias, da mesma forma como em [13]. A Tabela 1 apresenta as informações das bases pequenas e a Tabela 2 as informações das bases médias-grandes. A coluna #instâncias indica o número de instâncias da base, a coluna #atributos(Num/Nom) indica a quantidade de atributos, onde Num significa a quantidade de atributos numéricos e Nom a quantidade de atributos nominais, e a última coluna #classes indica a quantidade de classes da base. Estas bases foram obtidas do repositório *Keel* [2], disponível em http://sci2s.ugr.es/keel/datasets.php.

Nos experimentos foi utilizada a abordagem *ten-fold cross validation*. Além disso, cada *fold* foi testado três vezes, logo, para cada base de dados foram efetuadas 30 execuções. Os resultados apresentados expressam a média destas 30 execuções.

Tabela 1. Informações das bases pequenas.

Base de dados	#instâncias	#atributos (Num/Nom)	#classes
australian	690	14 (8/6)	2
automobile	205	25 (15/10)	6
balance	625	4 (4/0)	3
bupa	345	6 (6/0)	2
cleveland	297	13 (13/0)	5
contraceptive	1473	9 (9/0)	3
crx	653	15 (6/9)	2
ecoli	336	7 (7/0)	8
german	1000	20 (7/13)	2
glass	214	9 (9/0)	7
haberman	306	3 (3/0)	2
heart	270	13 (13/0)	2
hepatitis	155	19 (19/0)	2
iris	150	4 (4/0)	3
newthyroid	215	5 (5/0)	3
pima	768	8 (8/0)	2
tae	151	5 (5/0)	3
vehicle	846	18 (18/0)	4
wine	178	13 (13/0)	3
wisconsin	699	9 (9/0)	2

Tabela 2. Informações das bases médias e grandes.

Base de dados	#instâncias	#atributos (Num/Nom)	#classes
abalone	4174	8 (7/1)	28
banana	5300	2 (2/0)	2
coil2000	9822	85 (85/0)	2
magic	19020	10 (10/0)	2
marketing	6876	13 (13/0)	9
page-blocks	5472	10 (10/0)	5
penbased	10992	16 (16/0)	10
phoneme	5404	5 (5/0)	2
ring	7400	20 (20/0)	2
satimage	6435	36 (36/0)	7
segment	2310	19 (19/0)	7
spambase	4597	57 (57/0)	2
texture	5500	40 (40/0)	11
thyroid	7200	21 (21/0)	3
titanic	2201	3 (3/0)	2
twonorm	7400	20 (20/0)	2

5.2 Configuração dos Experimentos

Para analisar os resultados, serão considerados dois cenários: Cenário 1 para bases "pequenas" e Cenário 2 para bases "médias-grandes"; sempre utilizando o K-NN como algoritmo de classificação. Em cada cenário, serão realizados dois experimentos: (1) classificação da base de dados sem seleção de instâncias (ou seja, considerando a base completa); e (2) no Cenário 1, comparar o desempenho em selecionar instâncias dos algoritmos NSGA-II, SPEA-II e RNG; e no Cenário 2, comparar o desempenho em selecionar instâncias dos algoritmos NSGA-II, SPEA-II e PBIL. A escolha dos algoritmos RNG para as bases pequenas e PBIL para as bases médias e grandes foi baseada nas conclusões do trabalho de [13], já comentado na Seção 2. Vale destacar que os resultados apresentados neste trabalho para os algoritmos RNG e PBIL são oriundos da execução destes, dado que os códigos-fonte estão disponíveis no *Keel Software Tool*, disponível em http://sci2s.ugr.es/keel/download.php.

Os testes foram realizados em uma máquina com processador Core i7-3770 3.4GHz, com 8Gb RAM. Para todas as tabelas de resultados que serão apresentadas nas próximas seções, haverão colunas identificadas com (avg) e (dev), onde (avg) indica que os resultados são a média das 30 execuções realizadas para cada base de dados, e (dev) indica o desvio padrão dos 30 resultados obtidos.

5.3 Cenário 1

Neste cenário foram realizados os experimentos com as bases de dados pequenas. A Tabela 3 apresenta a classificação destas bases, considerando todas as instâncias. A coluna Tempo está relacionada ao tempo de execução do classificador K-NN, expresso em minutos.

Tabela 3. Classificação de bases pequenas sem seleção de instâncias.

Base de dados	Acurácia (avg)	Acurácia (dev)	Tempo (avg)	Tempo (dev)
australian	0,6145	0,0824	0,000027	0,000084
automobile	0,3716	0,1400	0,000000	0,000000
balance	0,7904	0,0657	0,000008	0,000019
bupa	0,6410	0,0499	0,000002	0,000006
cleveland	0,3912	0,0799	0,000001	0,000004
contraceptive	0,4657	0,0461	0,000014	0,000010
crx	0,6220	0,0492	0,000004	0,000009
ecoli	0,8070	0,0663	0,000001	0,000003
german	0,6120	0,0301	0,000009	0,000012
glass	0,7626	0,1071	0,000000	0,000000
haberman	0,6958	0,0504	0,000001	0,000003
heart	0,6444	0,0482	0,000001	0,000003
hepatitis	0,8222	0,1226	0,000000	0,000000
iris	0,9600	0,0332	0,000000	0,000000
newthyroid	0,9626	0,0516	0,000000	0,000000
pima	0,6680	0,0588	0,000001	0,000004
tae	0,6242	0,1440	0,000001	0,000003
vehicle	0,6476	0,0428	0,000008	0,000014
wine	0,7699	0,0633	0,000001	0,000003
wisconsin	0,9578	0,0282	0,000003	0,000006
Média	**0,6915**	**0,0680**	**0,000004**	**0,000009**

A melhor acurácia na classificação foi encontrada para a base *newthyroid* e a pior para a base *automobile*, ambas destacadas na Tabela 3.

Na Tabela 4, utilizando o NSGA-II para selecionar instâncias, pode-se verificar uma média de redução do número de instâncias das bases de dados de 52,82%. Comparando-se a média da acurácia na fase de testes entre a Tabela 3 e a Tabela 4, percebe-se uma pequena redução, de 0,6915 para 0,6726.

Usando o SPEA-II na seleção de instâncias, pode-se verificar uma média de redução do número de instâncias das bases de dados de 52,63% (Tabela 5). Comparando-se a média da acurácia na fase de testes entre a Tabela 3 e a Tabela 5, percebe-se também uma pequena redução, de 0,6915 para 0,6687.

A coluna Tempo nas Tabelas 4, 5 e 7 está relacionada ao tempo de execução empregado na seleção de instâncias, descrito em minutos.

A Tabela 6 descreve todos os parâmetros usados na seleção de instâncias pelo NSGA-II e pelo SPEA-II. Tais parâmetros foram definidos de forma empírica, com experimentos combinando diferentes valores para cada parâmetro. Os valores testados para o tamanho da população foram 50, 100, 150 e 200. Para o número máximo de avaliações foram considerados os valores 1000, 2000 e 3000. A taxa de probabilidade de cruzamento foi avaliada com os valores 0,5, 0,7 e 0,9. Por fim, os valores testados para a taxa de probabilidade de mutação foram 0,2, 0,4 e 0,6.

Segundo as conclusões do trabalho de [13], o algoritmo *Relative Neighborhood Graph Editing* (RNG) é o mais indicado para bases de dados pequenas, dentre 36 algoritmos de seleção de instâncias aplicados em 37 bases de dados de diferentes tamanhos. Desta forma, o algoritmo RNG também foi utilizado para comparação de resultados.

Tabela 4. Classificação de bases pequenas usando NSGA-II para selecionar instâncias.

Base de Dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia no Treinamento (avg)	Acurácia no Treinamento (dev)	Acurácia no Teste (avg)	Acurácia no Teste (dev)
australian	0,0480	0,0099	0,5220	0,0097	0,3246	0,0082	0,6188	0,0505
automobile	0,0017	0,0004	0,5653	0,0239	0,2408	0,0213	0,3598	0,1246
balance	0,0284	0,0042	0,5161	0,0115	0,4100	0,0115	0,8106	0,0513
bupa	0,0046	0,0004	0,5314	0,0154	0,3377	0,0145	0,6251	0,0635
cleveland	0,0035	0,0004	0,5443	0,0174	0,2413	0,0134	0,3988	0,0722
contraceptive	0,2435	0,0338	0,5162	0,0080	0,2490	0,0065	0,4725	0,0366
crx	0,0383	0,0062	0,5234	0,0106	0,3224	0,0090	0,6270	0,0679
ecoli	0,0045	0,0005	0,5227	0,0149	0,4160	0,0150	0,7917	0,0770
german	0,1088	0,0099	0,5174	0,0093	0,3228	0,0090	0,6120	0,0404
glass	0,0020	0,0002	0,5339	0,0163	0,3968	0,0148	0,6785	0,1166
haberman	0,0036	0,0004	0,5289	0,0152	0,3729	0,0128	0,7023	0,0591
heart	0,0029	0,0004	0,5347	0,0128	0,3547	0,0127	0,6580	0,0806
hepatitis	0,0007	0,0002	0,5496	0,0207	0,4398	0,0221	0,7742	0,1428
iris	0,0013	0,0002	0,5121	0,0136	0,4879	0,0136	0,9622	0,0379
newthyroid	0,0020	0,0002	0,5158	0,0118	0,4813	0,0121	0,9147	0,0445
pima	0,0519	0,0073	0,5197	0,0087	0,3562	0,0091	0,6785	0,0536
tae	0,0013	0,0002	0,5492	0,0188	0,3427	0,0201	0,4586	0,1304
vehicle	0,0643	0,0088	0,5155	0,0092	0,3436	0,0073	0,6283	0,0621
wine	0,0015	0,0003	0,5399	0,0173	0,4078	0,0168	0,7212	0,1110
wisconsin	0,0369	0,0056	0,5047	0,0083	0,4871	0,0081	0,9593	0,0309
Média	**0,0325**	**0,0045**	**0,5282**	**0,0137**	**0,3668**	**0,0129**	**0,6726**	**0,0727**

Tabela 5. Classificação de bases pequenas usando SPEA-II para selecionar instâncias.

Base de Dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia no Treinamento (avg)	Acurácia no Treinamento (dev)	Acurácia no Teste (avg)	Acurácia no Teste (dev)
australian	0,0417	0,0074	0,5225	0,0108	0,3208	0,0084	0,6319	0,0770
automobile	0,0025	0,0005	0,5638	0,0230	0,2352	0,0229	0,3242	0,1431
balance	0,0292	0,0039	0,5127	0,0085	0,4105	0,0088	0,8000	0,0432
bupa	0,0053	0,0006	0,5301	0,0129	0,3341	0,0099	0,6300	0,0603
cleveland	0,0044	0,0006	0,5376	0,0189	0,2389	0,0143	0,4212	0,0756
contraceptive	0,2415	0,0309	0,5161	0,0078	0,2477	0,0073	0,4695	0,0361
crx	0,0352	0,0062	0,5227	0,0096	0,3196	0,0092	0,6360	0,0538
ecoli	0,0051	0,0006	0,5230	0,0119	0,4127	0,0119	0,7898	0,0549
german	0,1050	0,0125	0,5159	0,0077	0,3207	0,0085	0,6090	0,0344
glass	0,0027	0,0003	0,5364	0,0149	0,3903	0,0137	0,6723	0,1353
haberman	0,0045	0,0006	0,5254	0,0116	0,3712	0,0129	0,6763	0,0623
heart	0,0036	0,0004	0,5361	0,0149	0,3494	0,0129	0,6346	0,0901
hepatitis	0,0012	0,0002	0,5449	0,0228	0,4398	0,0246	0,7751	0,1586
iris	0,0025	0,0006	0,5136	0,0109	0,4864	0,0109	0,9667	0,0420
newthyroid	0,0028	0,0003	0,5115	0,0122	0,4842	0,0130	0,9193	0,0490
pima	0,0527	0,0057	0,5158	0,0077	0,3548	0,0103	0,6715	0,0581
tae	0,0020	0,0003	0,5421	0,0189	0,3432	0,0233	0,4475	0,1311
vehicle	0,0639	0,0078	0,5172	0,0062	0,3408	0,0050	0,6252	0,0449
wine	0,0024	0,0004	0,5306	0,0163	0,4076	0,0142	0,7120	0,0946
wisconsin	0,0363	0,0055	0,5072	0,0063	0,4840	0,0073	0,9618	0,0309
Média	**0,0322**	**0,0043**	**0,5263**	**0,0127**	**0,3646**	**0,0125**	**0,6687**	**0,0738**

Tabela 6. Parâmetros do NSGA-II e do SPEA-II para bases pequenas.

Parâmetro	Valor
Tamanho da população	50
Número máximo de avaliações	1000
Probabilidade de cruzamento	0,9
Probabilidade de mutação	0,2

Na Tabela 7, utilizando o algoritmo RNG, pode-se observar uma média de redução de 24,90% no número de instâncias. Comparando-se a média da acurácia na fase de testes entre a Tabela 3 e a Tabela 7, percebe-se uma pequena variação positiva, de 0,6915 para 0,7061.

Comparando-se a abordagem sem seleção de instâncias com os algoritmos NSGA-II, SPEA-II e RNG, é possível observar similaridade na média de acurácia, porém reduzindo pela metade o número de instâncias, no caso do NSGA-II e do SPEA-II, e em aproximadamente 25% com o RNG, o que contribui significativa e positivamente na redução de tempo de execução do K-NN. Isso porque, como o K-NN é um algoritmo determinístico, quanto menor o número de instâncias a serem comparadas, mais rápido é o processo de classificação. Apesar da taxa de redução do número

de instâncias do RNG ser inferior, seu tempo gasto para selecionar as instâncias foi menor em quase todas as bases, resultando em uma redução de tempo média de 58% em relação ao NSGA-II e ao SPEA-II.

Tabela 7. Classificação de bases pequenas usando RNG para selecionar instâncias.

Base de dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia (avg)	Acurácia (dev)
australian	0,0142	0,0009	0,1599	0,0093	0,6464	0,0755
automobile	0,0013	0,0001	0,3328	0,0232	0,3796	0,1368
balance	0,0346	0,0017	0,1284	0,0071	0,8576	0,0401
bupa	0,0047	0,0005	0,3855	0,0195	0,6232	0,0594
cleveland	0,0033	0,0006	0,4224	0,0099	0,5075	0,0755
contraceptive	0,0722	0,0024	0,5019	0,0075	0,5201	0,0366
crx	0,0191	0,0027	0,1513	0,0094	0,6509	0,0534
ecoli	0,0040	0,0001	0,2133	0,0095	0,8307	0,0567
german	0,0403	0,0023	0,279	0,0076	0,6680	0,0264
glass	0,0018	0,0002	0,3349	0,0131	0,6952	0,1193
haberman	0,0021	0,0001	0,358	0,0203	0,7281	0,0570
heart	0,0024	0,0001	0,2119	0,0171	0,6704	0,0662
hepatitis	0,0007	0,0001	0,1835	0,0262	0,8338	0,1279
iris	0,0015	0,0001	0,0489	0,0102	0,9600	0,0332
newthyroid	0,0015	0,0001	0,0517	0,0074	0,9258	0,0685
pima	0,0169	0,0017	0,2776	0,0107	0,7371	0,0559
tae	0,0012	0,0001	0,5467	0,0294	0,5113	0,1354
vehicle	0,0222	0,0019	0,2972	0,0101	0,6512	0,0425
wine	0,0020	0,0003	0,0599	0,0139	0,7588	0,0608
wisconsin	0,0211	0,0006	0,0348	0,0036	0,9666	0,0276
Média	**0,0134**	**0,0008**	**0,2490**	**0,0133**	**0,7061**	**0,0677**

A Tabela 8 apresenta o resumo das médias dos resultados dos experimentos.

Tabela 8. Comparação dos resultados das bases pequenas.

Abordagem	Tempo	Taxa de redução	Acurácia
Sem seleção de instâncias	-	-	0,6915
NSGA-II	0,0325	0,5282	0,6726
SPEA-II	0,0322	0,5263	0,6687
RNG	0,0134	0,2490	0,7061

5.4 Cenário 2

No Cenário 2 são consideradas as bases de dados médias e grandes. A Tabela 9 apresenta os resultados da classificação destas bases sem seleção de instâncias.

Tabela 9. Classificação de bases médias e grandes sem seleção de instâncias.

Bases de dados	Acurácia (avg)	Acurácia (dev)	Tempo (avg)	Tempo (dev)
abalone	0,2065	0,0154	0,000060	0,000111
banana	0,8728	0,0115	0,000009	0,000049
coil2000	0,8991	0,0100	0,000026	0,000078
magic	0,7832	0,0088	0,000221	0,000467
marketing	0,2842	0,0132	0,000051	0,000103
page-blocks	0,9580	0,0074	0,000017	0,000066
penbased	0,9935	0,0022	0,000079	0,000123
phoneme	0,9051	0,0167	0,000018	0,000068
ring	0,7539	0,0086	0,000043	0,000097
satimage	0,9058	0,0123	0,000043	0,000097
segment	0,9671	0,0184	0,000000	0,000000
spambase	0,8216	0,0251	0,000008	0,000046
texture	0,9909	0,0030	0,000008	0,000046
thyroid	0,9299	0,0078	0,000017	0,000063
titanic	0,5916	0,0780	0,000000	0,000000
twonorm	0,9465	0,0068	0,000034	0,000089
Média	**0,7909**	**0,0153**	**0,000038**	**0,000091**

A melhor acurácia na classificação foi encontrada para a base *penbased*, no valor médio de 0,9935, e a pior para a base *abalone*, com valor médio igual a 0,2065, ambas destacadas na Tabela 9. A coluna Tempo está relacionada ao tempo de execução do classificador K-NN, expresso em minutos.

Utilizando o NSGA-II para selecionar instâncias nas bases médias e grandes, pode-se verificar uma média de redução de 50,38% do número de instâncias das bases de dados, conforme pode ser verificado na Tabela 10. Comparando-se a média da acurácia na fase de testes entre a Tabela 9 e a Tabela 10, percebe-se uma pequena melhora, de 0,7909 para 0,7921, respectivamente.

De acordo com os resultados mostrados na Tabela 11, usando o SPEA-II na seleção de instâncias, pode-se verificar uma média de redução do número de instâncias das bases de dados de 50,4%. Observando-se a média da acurácia na fase de testes entre a Tabela 9 e a Tabela 11, é possível perceber uma pequena melhora, de 0,7909 para 0,7931.

Tabela 10. Classificação de bases médias e grandes usando NSGA-II para selecionar instâncias.

Base de Dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia no Treinamento (avg)	Acurácia no Treinamento (dev)	Acurácia no Teste (avg)	Acurácia no Teste (dev)
abalone	2,6009	0,2684	0,5121	0,0047	0,1101	0,0020	0,2094	0,0159
banana	4,6786	0,7076	0,5034	0,0031	0,4383	0,0031	0,8625	0,0135
coil2000	10,6461	0,8236	0,5023	0,0020	0,4520	0,0020	0,9003	0,0069
magic	35,1591	0,7228	0,5028	0,0014	0,3941	0,0013	0,7696	0,0101
marketing	9,8895	1,8133	0,5084	0,0044	0,1505	0,0029	0,2797	0,0173
page-blocks	5,1013	0,7764	0,5012	0,0021	0,4813	0,0021	0,9544	0,0077
penbased	10,7884	0,2320	0,5008	0,0015	0,4969	0,0015	0,9915	0,0028
phoneme	4,9035	0,7559	0,5017	0,0020	0,4580	0,0024	0,8740	0,0135
ring	11,6837	1,9386	0,5044	0,0026	0,3800	0,0027	0,7209	0,0099
satimage	7,5134	1,5099	0,5032	0,0021	0,4563	0,0020	0,8947	0,0134
segment	0,6597	0,1097	0,5025	0,0032	0,4850	0,0036	0,9453	0,0215
spambase	3,0294	0,4293	0,5043	0,0035	0,4175	0,0036	0,7887	0,0168
texture	4,9011	0,8333	0,5002	0,0025	0,4968	0,0024	0,9825	0,0041
thyroid	10,5249	2,0377	0,5017	0,0021	0,4671	0,0024	0,9211	0,0089
titanic	0,5723	0,0780	0,5107	0,0058	0,3072	0,0329	0,6318	0,1181
twonorm	11,5908	1,9573	0,5017	0,0023	0,4754	0,0024	0,9464	0,0094
Média	**8,3902**	**0,9371**	**0,5038**	**0,0028**	**0,4042**	**0,0043**	**0,7921**	**0,0181**

Tabela 11. Classificação de bases médias e grandes usando SPEA-II para selecionar instâncias.

Base de Dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia no Treinamento (avg)	Acurácia no Treinamento (dev)	Acurácia no Teste (avg)	Acurácia no Teste (dev)
abalone	2,5844	0,2269	0,5109	0,0048	0,1096	0,0022	0,2095	0,0157
banana	4,6783	0,7112	0,5040	0,0022	0,4379	0,0019	0,8660	0,0120
coil2000	12,3055	1,6635	0,5023	0,0021	0,4516	0,0023	0,9008	0,0097
magic	36,0893	0,7967	0,5027	0,0019	0,3938	0,0016	0,7734	0,0098
marketing	9,6979	1,7119	0,5088	0,0033	0,1496	0,0019	0,2790	0,0146
page-blocks	4,8488	0,7990	0,5021	0,0020	0,4801	0,0021	0,9548	0,0067
penbased	18,0015	2,3949	0,5000	0,0013	0,4978	0,0013	0,9914	0,0028
phoneme	3,0685	0,2548	0,5030	0,0022	0,4566	0,0028	0,8771	0,0149
ring	6,1478	0,4555	0,5045	0,0028	0,3802	0,0029	0,7226	0,0107
satimage	4,3390	0,1409	0,5031	0,0022	0,4560	0,0028	0,8952	0,0117
segment	0,4195	0,0355	0,5023	0,0030	0,4851	0,0033	0,9489	0,0164
spambase	3,1699	0,4157	0,5051	0,0030	0,4164	0,0032	0,7907	0,0194
texture	5,2674	0,8716	0,5004	0,0021	0,4966	0,0022	0,9828	0,0057
thyroid	12,2846	2,4873	0,5022	0,0019	0,4664	0,0021	0,9208	0,0060
titanic	0,6982	0,1395	0,5111	0,0049	0,3055	0,0315	0,6329	0,1281
twonorm	12,4232	2,4752	0,5016	0,0021	0,4754	0,0021	0,9445	0,0087
Média	**8,5015**	**0,9738**	**0,5040**	**0,0026**	**0,4037**	**0,0041**	**0,7931**	**0,0183**

A coluna Tempo nas Tabelas 10, 11 e 13 está relacionada ao tempo de execução empregado na seleção de instâncias, descrito em minutos.

Os parâmetros usados na seleção de instâncias pelo NSGA-II e pelo SPEA-II para bases médias e grandes foram definidos empiricamente, assim como feito para as bases pequenas, e são apresentados na Tabela 12.

Tabela 12. Parâmetros do NSGA-II e do SPEA-II para bases médias e grandes.

Parâmetro	Valor
Tamanho da população	100
Número máximo de avaliações	1000
Probabilidade de cruzamento	0,9
Probabilidade de mutação	0,2

Também de acordo com as conclusões do trabalho de [13], o algoritmo *Populational Based Incremental Learning* (PBIL) é o mais indicado para as bases médias e grandes. Assim, os resultados do algoritmo PBIL foram utilizados para comparação.

Tabela 13. Classificação de bases médias e grandes usando PBIL para selecionar instâncias.

Base de dados	Tempo (avg)	Tempo (dev)	Taxa de Redução (avg)	Taxa de Redução (dev)	Acurácia (avg)	Acurácia (dev)
abalone	7,6955	0,3357	0,8920	0,0045	0,2384	0,0167
banana	8,5233	0,6939	0,8997	0,0067	0,8904	0,0097
coil2000	238,3983	18,5149	0,8815	0,0029	0,9321	0,0048
magic	214,4137	4,4082	0,8553	0,0019	0,7666	0,0116
marketing	28,5009	0,9345	0,8752	0,0026	0,2931	0,0107
page-blocks	13,9045	0,3070	0,9113	0,0018	0,9421	0,0065
penbased	87,5008	1,3720	0,8893	0,0021	0,9864	0,0043
phoneme	11,1851	0,2080	0,8877	0,0029	0,8642	0,0136
ring	47,7867	0,6475	0,8667	0,0016	0,7850	0,0115
satimage	44,6782	0,5406	0,8951	0,0015	0,8869	0,0134
segment	2,9151	0,0547	0,9373	0,0036	0,8411	0,0339
spambase	31,0191	0,6299	0,8977	0,0051	0,7268	0,0254
texture	32,6490	0,4144	0,9073	0,0026	0,9656	0,0083
thyroid	42,8410	0,7282	0,8995	0,0033	0,9328	0,0057
titanic	1,3415	0,0285	0,9639	0,0015	0,7515	0,0316
twonorm	44,1937	0,5339	0,8958	0,0023	0,9535	0,0094
Média	**53,5967**	**1,8970**	**0,8972**	**0,0029**	**0,7973**	**0,0136**

Na Tabela 13, utilizando o algoritmo PBIL, pode-se observar uma média de redução de 89,72% no número de instâncias. Comparando-se a média da acurácia na fase de testes entre a Tabela 9 e a Tabela 13, percebe-se uma pequena variação positiva, de 0,7909 para 0,7973.

Ao se comparar a abordagem sem seleção de instâncias com as abordagens utilizando os algoritmos NSGA-II, SPEA-II e PBIL, é possível observar similaridade na média de acurácia entre elas. Entretanto, o algoritmo PBIL apresentou uma taxa média de redução do número de instâncias de 89,72%, bastante superior aos algoritmos NSGA-II e SPEA-II, cujas taxas de redução estiveram em torno dos 50%.

Por fim, o tempo gasto para selecionar as instâncias foi menor em todas as bases para os algoritmos NSGA-II e SPEA-II, com valores de tempo médios de 8,3902 minutos e 8,5015 minutos, respectivamente. O algoritmo PBIL teve um valor médio de tempo de 53,5967 minutos, aproximadamente seis vezes maior que os tempos médios dos algoritmos NSGA-II e SPEA-II. A Tabela 14 apresenta o resumo das médias dos resultados dos experimentos realizados.

Tabela 14. Comparação dos resultados das bases médias e grandes.

Abordagem	Tempo	Taxa de redução	Acurácia
Sem seleção de instâncias	-	-	0,7909
NSGA-II	8,3902	0,5038	0,7921
SPEA-II	8,5015	0,5040	0,7931
PBIL	53,5967	0,8972	0,7973

6. CONCLUSÕES

Este trabalho teve como objetivo avaliar o desempenho dos algoritmos evolutivos multiobjetivo NSGA-II e SPEA-II em realizar seleção de instâncias para problemas de classificação, utilizando como classificador o algoritmo K-NN. Os experimentos foram realizados com 36 bases de dados de tamanhos diferentes e os resultados foram comparados com os obtidos sem realizar a seleção de instâncias (com as bases de dados completas) e com outros resultados apresentados nos trabalhos correlatos disponíveis na literatura. Foram avaliados o tempo de execução dos algoritmos aplicados na seleção de instâncias, a taxa de redução do número de instâncias apresentada por cada algoritmo e a acurácia nos resultados dos testes.

Em um primeiro cenário, foram realizados dois experimentos para bases consideradas pequenas, avaliando primeiro a classificação da base de dados sem seleção de instâncias; e posteriormente, realizando a comparação dos algoritmos NSGA-II, SPEA-II e RNG no processo de selecionar instâncias, com todos os experimentos utilizando o K-NN como classificador.

Através dos resultados, foi possível observar que os valores de acurácia foram similares entre as quatro abordagens. Com relação ao tempo de execução no processo de seleção de instâncias, observa-se que os três algoritmos tiveram tempos de resposta reduzidos, sendo que o algoritmo RNG apresentou o menor tempo de execução, inferior em 58% em relação ao NSGA-II e ao SPEA-II. Por fim, analisando a taxa de redução, os algoritmos NSGA-II e SPEA-II se mostraram superiores ao RNG, reduzindo em torno de 53% a quantidade de instâncias das bases de dados.

Com valores similares de acurácia e tempos de execução pequenos, a diferença entre as abordagens se dá pela redução do número de instâncias, destacando os algoritmos NSGA-II e SPEA-II como os mais adequados para as bases pequenas.

No segundo cenário, foram realizados dois experimentos para bases consideradas médias e grandes, similares aos realizados no Cenário 1, avaliando primeiro a classificação da base de dados sem seleção de instâncias; e depois comparando os algoritmos NSGA-II, SPEA-II e PBIL no processo de seleção de instâncias, também utilizando o K-NN como classificador.

É possível verificar, com base nos resultados, que os valores de acurácia foram bastante similares entre as quatro abordagens. Já no caso do tempo de execução no processo de seleção de instâncias, os algoritmos NSGA-II e SPEA-II tiveram tempos de resposta bastante inferiores ao tempo do PBIL. Entretanto, analisando a taxa de redução, o algoritmo PBIL se mostrou superior ao NSGA-II e ao SPEA-II, reduzindo em torno de 90% a quantidade de instâncias das bases de dados.

Desta forma, algoritmo PBIL se mostrou mais adequado que os algoritmos NSGA-II e SPEA-II para bases médias e grandes, por obter uma elevada taxa de redução do número de instâncias sem apresentar alterações significativas nos valores de acurácia. No entanto, como o tempo de execução do algoritmo PBIL é aproximadamente seis vezes maior que os tempos do NSGA-II e do SPEA-II, sua aplicabilidade deve ser avaliada em cada problema a ser tratado.

Enfim, independente do algoritmo evolutivo multiobjetivo utilizado para selecionar instâncias, os resultados demostraram a eficiência tanto do NSGA-II quanto do SPEA-II em obter um subconjunto de dados com tamanho menor e com um desempenho

para classificação similar ou até mesmo maior que o original, por eliminar redundâncias e ruídos, contribuindo significativamente na redução do tempo de execução, dos requisitos de memória e da sensibilidade a ruídos do classificador.

7. REFERÊNCIAS

[1] Aha, D., Kibler, D. and Albert, M. 1991. Instance-based learning algorithms. *Machine Learning*, vol.6, n.1, 37-66.

[2] Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S., Sánchez, L. e Herrera, F. 2011. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, vol.17, n.2-3, 255-287.

[3] Antonelli, M., Ducange, P. and Marcelloni, F. 2012. Genetic Training Instance Selection in Multiobjective Evolutionary Fuzzy Systems: A Coevolutionary Approach. *IEEE Transactions on Fuzzy Systems*, vol.20, n.2, 276-290.

[4] Baluja, S. 1994. *Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning*, Carnegie Mellon University, Pittsburgh - PA - USA.

[5] Brighton, H. and Mellish, C. 2002. Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery*, vol.6, n.2, 153-172.

[6] Cano, J. R., Herrera, F. and Lozano, M. 2003. Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. *IEEE Transactions on Evolutionary Computation*, vol.7, n.6, 561-575.

[7] Chen, J.-H., Chen, H.-M. and Ho, S.-Y. 2005. Design of nearest neighbor classifiers: multi-objective approach. *International Journal of Approximate Reasoning*, vol.40, n.1–2, 3-22.

[8] Coello, C. A. C. 2001. A Short Tutorial on Evolutionary Multiobjective Optimization. In *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*, Springer-Verlag, 21-40.

[9] Coello, C. A. C., Lamont, G. B. and Veldhuizen, D. A. V. 2007. *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2. ed., Springer, New York.

[10] Deb, K. 2001. *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, John Wiley & Sons.

[11] Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, vol.6, n.2, 182-197.

[12] Eshelman, L. J. 1991. *The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination*. Foundations of Genetic Algorithms. G. J. E. Rawlings, Morgan Kaufmann, 265-283.

[13] Fazzolari, M., Giglio, B., Alcalá, R., Marcelloni, F. and Herrera, F. 2013. A study on the application of instance selection techniques in genetic fuzzy rule-based classification systems: Accuracy-complexity trade-off. *Knowledge-Based Systems*, vol.54, 32-41.

[14] Fernández, A., López, V., del Jesus, M. J. and Herrera, F. 2015. Revisiting Evolutionary Fuzzy Systems: Taxonomy, applications, new trends and challenges. *Knowledge-Based Systems*, vol.80, 109-121.

[15] García-Pedrajas, N., Haro-García, A., and Pérez-Rodríguez, J. 2013. A scalable approach to simultaneous evolutionary instance and feature selection. *Information Sciences,* vol. 228, 150-174.

[16] García-Pedrajas, N. and J. Pérez-Rodríguez 2012. "Multi-selection of instances: A straightforward way to improve evolutionary instance selection." *Applied Soft Computing*, vol. 12, n.11, 3590-3602.

[17] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley.

[18] Ho, S.-Y., Liu, C.-C., Liu, S. and Jou, J.-W. 2002. Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. *Pattern Recognition Letters*, vol.23, n.13, 1495–1503.

[19] Liu, H. and H. Motoda 2002. On Issues of Instance Selection. *Data Mining and Knowledge Discovery*, vol. 6, n. 2, 115-130.

[20] Miller, B. and D. Goldberg 1995. Genetic Algorithms, Tournament Selection, and the Effects of Noise. *Complex Systems*, vol. 9, 193–212.

[21] Tsai, C.-F. and Chen, Z.-Y. 2014. Towards high dimensional instance selection: An evolutionary approach. *Decision Support Systems*, vol. 61, 79-92.

[22] Tsai, C.-F., Chen, Z.-Y., and Ke, S.-W. 2014. Evolutionary instance selection for text classification. *The Journal of Systems and Software*, vol. 90, 104-113.

[23] Whitley, D. 1989. The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In *Proceedings of the Third International Conference on Genetic Algorithms*. George Mason University, USA, Morgan Kaufmann Publishers Inc., 116-121.

[24] Wilson, D. R. and T. Martinez 2000. Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning*, vol. 38, n. 3, 257-286.

[25] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J. and Steinberg, D. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems*, vol. 14, n. 1, 1-37.

[26] Zitzler, E. and L. Thiele 1999. Multiobjective evolutionary algorithms: a comparative case study and the Strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, vol. 3, n. 4, 257-271.

[27] Zitzler, E., Laumanns, M. and Thiele, L. 2001. *SPEA2: Improving the Strength Pareto Evolutionary Algorithm*, Technical Report, Swiss Federal Institute of Technology, Department of Electrical Engineering.