Instance Genetic Selection for Fuzzy Rule-Based Systems Optimization to Opinion Classification

T. Cerqueira, F. Bertoni and M. Pires

***Abstract*—Opinion Mining aims to identify people feelings about some item of interest based on content available on the Web, without the user having to find and read all the news about it. As opinions are related to feelings that are often described by imprecise terms, Fuzzy Systems appear as an alternative to treat the information subjectivity. An important task in developing Fuzzy Systems is define the rule base usually based on textual databases in case of opinion mining. However, these databases are extensive and the algorithms used to generate the rule base often result in many rules, making it difficult to achieve accuracy with a low computational cost. To deal with this problem, instance selection can be applied to reduce databases in order to save only relevant data. Thus, the aim of the present study is optimize a Fuzzy System, using instance selection to generate a reduced rule base. As the issue mentioned is a multiobjective problem, which seeks to increase accuracy and reduce the number of rules, we have chosen to apply a Multiobjective Genetic Algorithm, since it has been acknowledged as a promising approach in the literature. The results demonstrate that the Multiobjective Genetic Algorithms can be applied in instance selection for opinion classification problems, presenting a reduction in the number of instances and execution time, without significant changes in accuracy.**

***Index Terms*—Opinion mining, Fuzzy classification, Multiobjective genetic algorithms, Instance selection.**

I. Introdução

O aumento do número de pessoas com acesso à internet nos últimos anos proporcionou a geração de um grande volume de dados publicados por meio de páginas, fóruns e redes sociais. Por estes meios, os usuários podem expressar suas opiniões sobre diversos temas como produtos, serviços, filmes e política, e utilizar opiniões disponíveis na internet para tomar decisões sobre assuntos específicos. Contudo, é uma tarefa difícil e muitas vezes impraticável para o usuário humano encontrar e sumarizar informações relevantes [1]. Além do grande volume de dados gerados todos os dias em diferentes tipos de formatos, extrair opiniões de textos em linguagem natural pode ser uma tarefa complicada, pois muitas vezes sentimentos são descritos por termos subjetivos ou imprecisos. Outro aspecto importante é a dificuldade em definir o sentimento geral de uma sentença ou de um documento quando opiniões de graus diferentes se combinam, como por exemplo: "é um ótimo celular e tem um acabamento muito bom, mas a bateria é péssima"; "Este tênis é confortável, mas a durabilidade é ruim". Essas dificuldades para se determinar o sentimento ou a polaridade (positiva ou negativa) de uma sentença implicam na necessidade de desenvolvimento de eficientes sistemas de análise de sentimento automatizados [2] [3] [4]. Neste contexto, a Lógica Fuzzy, introduzida por Zadeh [5], surge como alternativa viável em função de sua característica de tratar variáveis semânticas com certo grau de imprecisão, sendo quase intuitiva a aproximação das regras de especificação de opiniões ou sentimentos, dos atributos de um Sistema Fuzzy.

O conhecimento sobre o domínio do problema é armazenado na base de regras de um Sistema Fuzzy. Esta base pode ser construída por um especialista ou gerada por algoritmos específicos, a partir das variáveis de entrada do sistema, como por exemplo, o algoritmo de Wang-Mendel [6]. Basicamente, o algoritmo de Wang-Mendel propõe a geração de uma regra para cada instância da base de dados, e em seguida, a remoção das regras conflitantes. Bases de dados utilizadas na mineração de opiniões geralmente apresentam uma quantidade muito grande de exemplos ou instâncias. Com isso, normalmente são geradas muitas regras, e algumas delas são redundantes ou não contribuem de maneira satisfatória no processo de classificação. Desta forma, técnicas para redução de bases de dados vêm sendo propostas. Essa redução pode ser alcançada de diversas maneiras, destacando-se a Seleção de Instâncias como um dos mecanismos mais promissores [7] [8]. Selecionar instâncias consiste em obter um representativo subconjunto de dados com tamanho menor e desempenho para classificação similar ou até mesmo maior que o original, dada a eliminação de instâncias supérfluas ou ruidosas.

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 de acurácia pretendida [9], caracterizando-se assim como um problema multiobjetivo, com objetivos conflitantes. Isso porque, ao reduzir o número de instâncias, pode-se retirar dados que seriam importantes para o processo de classificação, prejudicando a acurácia. Na busca do melhor equilíbrio entre redução dos dados e acurácia, diversas metodologias vêm sendo propostas para selecionar instâncias, e dentre elas os Algoritmos Genéticos Multiobjetivo (AGMO).

Nesse contexto, este trabalho se propõe a avaliar a aplicação do AGMO NSGA-II *(Non-dominated Sorting Genetic Algorithm)* [10] na seleção de instâncias em bases de dados de opiniões, como mecanismo para otimização da base de regras de um Sistema Baseado em Regras Fuzzy, de forma a melhorar o desempenho do classificador, através da redução da base de regras, e por consequência do tempo de execução do sistema, mantendo ou até mesmo melhorando o poder de classificação,

Tayane L. Cerqueira is with OpenSEV Company, Feira de Santana, Bahia, Brazil, e-mail: engcomp.tayane@gmail.com.

Fabiana C. Bertoni and Matheus G. Pires are with Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil, e-mail: fcbertoni@gmail.com, mgpires@ecomp.uefs.br.

quando comparado ao sistema gerado a partir da base de dados original.

II. Trabalhos Relacionados

Esta seção apresenta alguns trabalhos que realizam seleção de instâncias para problemas de classificação, considerando diferentes tipos de classificadores e diferentes bases de dados. Não foram encontrados na literatura trabalhos que realizam seleção de instâncias para aplicações em Classificação de Opiniões, sendo este um aspecto de originalidade do trabalho.

No artigo desenvolvido por [11], foi abordado o problema de seleção de instâncias utilizando um algoritmo genético multi-objetivo com uma abordagem co-evolutiva, para o aprendizado de um Sistema Fuzzy. Foi utilizado um algoritmo genético para redução de instâncias do conjunto de treinamento. Um algoritmo evolutivo multiobjetivo foi aplicado na seleção de regras e escolha dos parâmetros dos conjuntos fuzzy, com o objetivo de aumentar a acurácia e minimizar a complexidade da base. Foram testadas 12 bases de dados cujo tamanho variou entre 4052 e 40768 instâncias. Os melhores resultados foram encontrados reduzindo o conjunto de treinamento de 10% a 20% do conjunto original, sendo que se obteve uma redução de 86,36% no tempo de execução.

O trabalho apresentado em [12] realizou um estudo para avaliar o desempenho das técnicas de seleção de instâncias e características baseado em algoritmos genéticos. Os autores realizaram experimentos de classificação com 4 bases de dados pequenas e 4 grandes utilizando a seleção de instâncias e de características de forma individual e associada. Ao realizar a classificação associando seleção de instâncias e características, foi observado uma pequena redução no desempenho quando comparado com a utilização das técnicas de forma individual. Entretanto, embora não tenha sido observada uma diferença significativa na acurácia, a combinação das duas técnicas reduziu grande parte do esforço computacional no treinamento dos classificadores. Considerando a eficácia e eficiência da classificação, os autores recomendam primeiramente a execução da seleção de características e, em seguida, a seleção de instâncias. Os classificadores utilizados no trabalho foram o SVM (*Support Vector Machine*) e KNN (*k-nearest neighbor*).

Os autores de [13] desenvolveram um trabalho de análise de desempenho entre 36 métodos diferentes de seleção de instâncias para sistemas de classificação genético-fuzzy baseado em regras. O objetivo foi investigar se os métodos seriam úteis para diminuir o tempo de execução e a complexidade da classificação. Neste trabalho, foram utilizadas 37 bases com tamanhos diferentes. Os autores concluíram que a escolha do método deve ser feita considerando o tamanho da base de dados. Para base de dados pequenas (entre 150 e 1473 instâncias), o método RNG (*Relative Neighborhood Graph Editing*) apresentou melhor desempenho. Para bases de dados de tamanho médio e grande (entre 2201 e 19020 instâncias), os métodos IGA (*Intelligent Genetic Algorithm*) e PBIL (*Populational Based Incremental Learning*) foram mais adequados por apresentarem um bom equilíbrio entre as taxas de acurácia e complexidade.

Em [14], os autores avaliaram o desempenho dos Algoritmos Multiobjetivo NSGA-II e SPEA-II para seleção de instâncias. Neste trabalho, foram utilizadas 36 bases de dados dentre pequenas, médias e grandes. A menor base de dados possuía 151 instâncias e a maior 10992 e o algoritmo de classificação utilizado foi o KNN (*k-nearest neighbor*). Em bases de dados pequenas, os algoritmos NSGA-II e SPEA-II foram comparados com o RNG (*Relative Neighborhood Graph Editing*), que apresentou o menor tempo de execução no processo de seleção de instâncias. Porém, as taxas de redução do NSGA-II e do SPEA-II foram superiores. Os valores de acurácia se mostraram similares. Em bases de dados grandes, o NSGA-II e o SPEA-II foram comparados com o PBIL (*Populational Based Incremental Learning*), que apresentou um tempo de execução aproximadamente seis vezes maior, mas, por outro lado, apresentou a maior taxa de redução. Por fim, os autores concluíram que 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 ao original.

III. Mineração de Opinião

A Mineração de Opinião, também chamada de Análise de Sentimento, é o campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos [2]. É muito comum, hoje em dia, as pessoas buscarem opiniões de outros usuários em *reviews*, sites de recomendação e discussões na *web*, para tomada de decisões antes de adquirir algum produto ou serviço, por exemplo. Além disso, as empresas não precisam mais realizar questionários e outros tipos de pesquisas para obter a opinião das pessoas, pois a informação que eles precisam é encontrada em abundância na *web*, em redes sociais, fóruns e *blogs*. Como boa parte dos dados encontrados estão na forma de texto, é necessária a utilização de técnicas capazes de extrair informações desse tipo de dado, técnicas estas de Mineração de Textos.

A Mineração de Textos é a descoberta, através de meios computacionais, de informações desconhecidas, utilizando-se ferramentas de extração automática de informação, a partir de documentos de textos não estruturados [15]. Contudo, nem sempre extrair opiniões de textos é uma tarefa fácil, pois, diferentemente de fatos, opiniões são subjetivas. Palavras que expressam opinião são utilizadas para mostrar sentimentos negativos ou positivos [1]. Por exemplo, “bom”, “legal” e “incrível” expressam sentimentos positivos, enquanto “ruim”, “horrível” e “péssimo” expressam sentimentos negativos. Entretanto, apesar das palavras de opinião serem importantes na análise de sentimento, elas não são suficientes. Em seu trabalho intitulado *”Sentiment analysis and opinion mining”*, Liu [2] relata os seguintes problemas: (1) uma palavra positiva ou negativa pode ter significados diferentes a depender do contexto em que está sendo empregada; (2) uma sentença contendo palavras de sentimento pode não expressar nenhum sentimento (esse tipo de sentença é comum em frases interrogativas, como por exemplo, “Você poderia me dizer se está

câmera é boa?"); (3) uso de ironias em sentenças que contêm palavras que expressam sentimento (exemplo: "O filme foi tão bom que dormi nos primeiros minutos"); (4) sentenças que não possuem nenhuma palavra de opinião, mas que apresentam uma opinião implícita, por exemplo: "Esta máquina de lavar utiliza uma grande quantidade de água".

IV. Sistemas de Classificação Baseados em Regras Fuzzy

Classificação é uma importante tarefa que é encontrada em várias áreas, tais como, reconhecimento de padrões, tomadas de decisão, mineração de dados e modelagem. A tarefa de classificação pode ser descrita da seguinte maneira: dado um conjunto de objetos $E = (e_1, e_2, ..., e_m)$, também chamados de padrões, exemplos ou instâncias, os quais são descritos por m características (ou atributos), designam uma classe c_i a partir de um conjunto de classes $C = (c_1, c_2, ..., c_j)$, para um exemplo e_p, o qual é descrito pelos valores de seus atributos $e_p = (a_{p1}, a_{p2}, ..., a_{pm})$.

Um Sistema de Classificação Baseado em Regras Fuzzy (SCBRF) é um Sistema Baseado em Regras Fuzzy (SBRF), onde as regras são projetadas para solucionar um problema de classificação. Estas regras possuem variáveis linguísticas em seus antecedentes, as quais definem as características do objeto, e no consequente de cada regra é definida uma classe. Uma regra fuzzy clássica possui a seguinte estrutura:

R_k : **SE** X_1 **é** A_{1l_1} **E** X_2 **é** A_{2l_2} **E** ... **E** X_m **é** A_{ml_T} **ENTÃO** $Classe = c_i$

onde R_k é o identificador da regra, $\mathbf{X} = (X_1, ..., X_m)$ são as características dos padrões, $P_m = (A_{ml_1}, ..., A_{ml_T})$ é a partição fuzzy da variável X_m, e $c_i \in C$ é a classe.

Em um SCBRF, o mecanismo de inferência aplica as regras fuzzy para um exemplo de entrada com o objetivo de determinar a classe a que ele pertence. O mecanismo de inferência Método de Raciocínio Fuzzy Geral (MRFG), proposto por [16], agrega os graus de compatibilidade entre as regras e classifica o padrão utilizando a classe que possui o maior grau de compatibilidade considerando todas as regras de mesma classe. O MRFG aplica os seguintes passos para classificar um exemplo e_p:

1) Calcular o grau de compatibilidade entre o padrão e_p e cada regra R_k, para $k = 1, ..., n$;
2) Calcular o valor de classificação $Classe_c$, para cada classe, onde $Classe_c$ é a agregação dos graus de compatibilidade de todas as regras com classe c_i, e representa a compatibilidade do padrão com todas as regras da classe c_i;
3) A classe com maior grau de classificação é atribuída ao padrão e_p.

V. Algoritmos Genéticos Multiobjetivo

Um Problema de Otimização Multiobjetivo (MOOP, do inglês *Multiobjective Optimization Problem*) possui um conjunto de funções objetivo a serem otimizadas (maximizadas ou minimizadas). No entanto, em geral, esses objetivos são conflitantes entre si, não permitindo que um objetivo possa ser melhorado sem comprometer os demais. 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 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. 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.

Os Algoritmos Genéticos Multiobjetivo (AGMO) 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 [17]. A aplicação dos AGMO para MOOP apresenta vantagens em relação às técnicas tradicionais, 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 AGMO [18]. Dentre os AGMO mais conhecidos destaca-se o NSGA-II (*Non-dominated Sorting Genetic Algorithm*), proposto por [10].

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 domina j, onde i pertence a 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.

VI. Seleção de Instâncias usando NSGA-II

Para o desenvolvimento desse trabalho, foi utilizada a base de dados criada por [19]. Neste trabalho, os autores realizaram o pré-processamento e extração de características da base de textos disponibilizada por Pang e Lee [20]. Esta base possui 2000 documentos com opiniões sobre filmes, previamente

classificados pelos autores, onde 1000 possuem opinião positiva e 1000 opinião negativa. Para transformar a base de textos em uma base de dados, com instâncias e características, os autores de [19] inicialmente classificaram cada palavra do texto utilizando sua classe gramatical. Cada documento de texto foi então representado por um vetor de palavras ou termos, chamados de *n-grams*, com suas respectivas classes. Em seguida, aplicaram técnicas para composição de *n-grams* e um algoritmo para atribuição de graus de polaridade (positiva ou negativa) para cada *n-gram* ou composição deles, conforme descrito em [21]. Finalmente, para formar as características, também baseado em [21], as técnicas de Soma, Contagem e Valor Máximo foram aplicadas. A técnica de Soma envolve a soma dos graus de polaridade para diferentes tipos de *n-grams*, como a soma das polaridades dos adjetivos de um documento. A Contagem é obtida de maneira semelhante para diferentes tipos de *n-grams*, contando o número de valores de polaridade positivos ou negativos, por exemplo. O Valor Máximo pode se referir a um máximo valor de polaridade de um *n-gram*. Como resultado, foram formadas 57 características.

A grande quantidade de regras gerada pelo algoritmo de Wang-Mendel, considerando a elevada quantidade de instâncias da base de dados para classificação de opinião, afeta diretamente o desempenho do SCBRF. Nesse contexto, foi utilizado o AGMO NSGA-II para selecionar instâncias que possuem maior poder de classificação, reduzindo a quantidade de regras geradas pelo algoritmo de Wang-Mendel, contribuindo para a redução do custo computacional do SCBRF, e ao mesmo tempo, mantendo a acurácia do classificador. O AGMO implementado possui as seguintes características:

- Os indivíduos são representados por um vetor binário de tamanho igual a quantidade de instâncias que serão selecionadas. Desta forma, cada instância da base é representada por um gene binário, onde o valor 0 indica a ausência da instância na base e o valor 1 indica sua presença;
- O tipo de cruzamento utilizado foi o *HUX crossover (Half Uniform Crossover)*, devido aos bons resultados obtidos por este operador nos trabalhos de [7], [22];
- A seleção dos cromossomos é feita pelo operador Torneio [23];
- O tipo de mutação utilizado foi o *Bit-Flip*.

A função de aptidão avalia os cromossomos a partir de dois objetivos: a taxa de acurácia e a taxa de redução da base de dados original, os quais são definidos pelas Equações 1 e 2, respectivamente.

$$taxa\ de\ acurácia = \frac{A}{N} \tag{1}$$

$$taxa\ de\ redução = \frac{(N - N')}{N} \tag{2}$$

A taxa de redução é calculada pela razão entre a quantidade de instâncias reduzidas (isto é, a diferença entre a quantidade total de instâncias (N) e a quantidade de instâncias selecionadas (N')) e a quantidade total de instâncias (N). O valor da variável A é a quantidade de acertos obtida por meio da aplicação do classificador KNN (*k-Nearest Neighbor*), com o valor de k igual a 1, por ser um algoritmo simples e com alto poder de classificação. A Tabela I descreve os parâmetros usados na seleção de instâncias utilizando o NSGA-II. Tais parâmetros foram definidos empiricamente, com experimentos combinando diferentes valores para cada parâmetro. Os valores testados para o tamanho da população inicial foram 50, 100, 150 e 200. Para o número máximo de avaliações foram considerados os valores 1000, 1500 e 2000. A taxa de probabilidade de cruzamento foi avaliada com os valores 60%, 80% e 90%. Por fim, os valores testados para a taxa de probabilidade de mutação foram 2% e 5%.

TABELA I
PARÂMETROS UTILIZADOS NO AGMO

Parâmetros	**Valor**
População inicial	10% da base de treinamento
Quantidade de avaliações	1000
Probabilidade de mutação	5%
Probabilidade de cruzamento	60%

TABELA II
EXECUÇÃO DO SCBRF UTILIZANDO A BASE DE DADOS ORIGINAL

	Quantidade de regras geradas	**Acurácia**	**Tempo de execução**
Fold 1	1255	60,0%	9s 85ms
Fold 2	1241	57,5%	8s 30ms
Fold 3	1259	64,0%	8s 26ms
Fold 4	1183	58,5%	8s 19ms
Fold 5	1182	59,5%	9s 40ms
Fold 6	1222	61,0%	8s 82ms
Fold 7	1255	56,5%	8s 62ms
Fold 8	1231	57,0%	9s 25ms
Fold 9	1200	55,0%	8s 77ms
Fold 10	1250	62,5%	11s 88ms
Média	**1227,8**	**59,1%**	**8s 27ms**

As variáveis de entrada do SCBRF são as 57 características da base de dados. Cada variável de entrada foi modelada com três conjuntos fuzzy triangulares distribuídos de maneira uniforme em relação ao domínio de valores de cada variável. Após a definição dos conjuntos fuzzy, foi utilizado o algoritmo de Wang-Mendel [6] para a geração da base de regras. Por fim, foi implementado o Método de Raciocínio Fuzzy Geral para classificar as instâncias.

VII. EXPERIMENTOS E RESULTADOS

Os experimentos para avaliação deste trabalho foram realizados por meio da abordagem *ten-fold cross validation*. Todos os experimentos foram executados em uma máquina com processador Intel Core i7-4510U com 2.00Ghz, 16GB RAM e sistema operacional Windows 10. Cada *fold* contém todas as instâncias da base de dados, onde, de forma aleatória, 90% da base é utilizada para a geração de regras (treinamento) e 10% para teste. Na primeira etapa dos experimentos, o classificador fuzzy é avaliado considerando a base de dados completa. Ao final da execução de cada *fold*, é calculadada a acurácia do classificador. Os resultados obtidos nesta primeira etapa são descritos na Tabela II.

O objetivo da segunda etapa dos experimentos é reduzir o número de regras do classificador fuzzy a partir da seleção

TABELA III
EXECUÇÃO DO SCBRF UTILIZANDO A BASE DE DADOS OTIMIZADA

	Instâncias Selecionadas	Regras geradas	Acurácia	Tempo de execução
Fold 1	651	607	58,0%	2s 17ms
Fold 2	667	611	56,5%	2s 12ms
Fold 3	667	632	60,0%	2s 98ms
Fold 4	655	614	63,5%	2s 57ms
Fold 5	661	613	59,0%	2s 36ms
Fold 6	678	642	65,0%	2s 64ms
Fold 7	668	620	59,5%	2s 57ms
Fold 8	656	613	63,0%	2s 33ms
Fold 9	660	629	57,0%	2s 65ms
Fold 10	660	628	62,0%	3s 00ms
Média	**662,3**	**620,9**	**60,3%**	**2s 53ms**

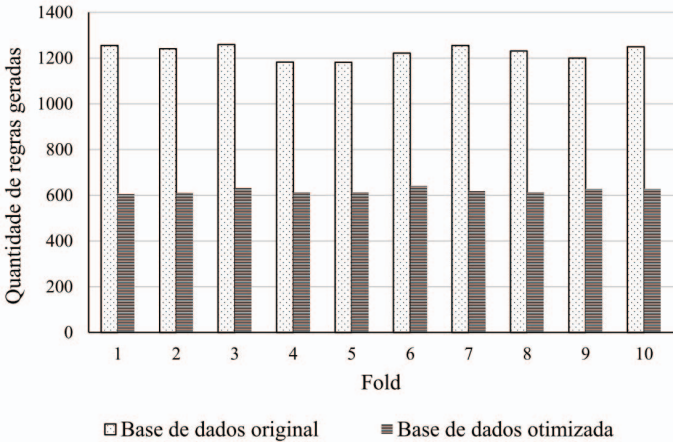

Fig. 1. Comparação entre a quantidade de regras geradas utilizando a base de dados original e base de dados otimizada.

genética das instâncias. Com um número menor de instâncias, o algoritmo de Wang-Mendel irá gerar uma base de regras menor, no entanto, será avaliado se esta redução afeta a acurácia na classificação. Os resultados são descritos na Tabela III.

Utilizando o AGMO NSGA-II para selecionar instâncias, foi obtida uma redução média de 66% no número de instâncias da base de dados original, conforme pode ser verificado nos dados mostrados na Tabela III em relação à quatidade total da base original, com 2000 instâncias. Além disso, utilizando a base de dados original foram geradas em média 1227,8 regras, enquanto que utilizando a base de dados otimizada, isto é, contendo as instâncias selecionadas pelo AGMO, foram geradas em média 620,9 regras. Esse valor corresponde a uma redução de aproximadamente 50,6% quando comparada a quantidade de regras geradas utilizando a base de dados original. A Figura 1 apresenta um gráfico comparativo entre as quantidades de regras geradas em cada *fold*, evidenciando de forma visual a significativa redução obtida.

Além da redução da base de regras, houve um pequeno aumento na acurácia do Sistema Fuzzy. Enquanto na utilização da base original a acurácia média foi de 59,1%, utilizando a base de dados otimizada pelo AGMO a acurácia aumentou para 60,3% em média. A comparação entre os resultados de acurácia obtidos em cada *fold* pode ser observada de forma visual na Figura 2. Apenas nos *folds* 1, 2, 3 e 5 a acurácia

TABELA IV
COMPARAÇÃO DOS RESULTADOS OBTIDOS UTILIZANDO A BASE DE DADOS ORIGINAL E A BASE DE DADOS OTIMIZADA

	Base original	Base otimizada
Acurácia média	59,1%	60,3%
Desvio padrão	0,028	0,028
Quantidade média de regras geradas	1227,8	620,9
Tempo médio de execução	8s 27ms	2s 53ms

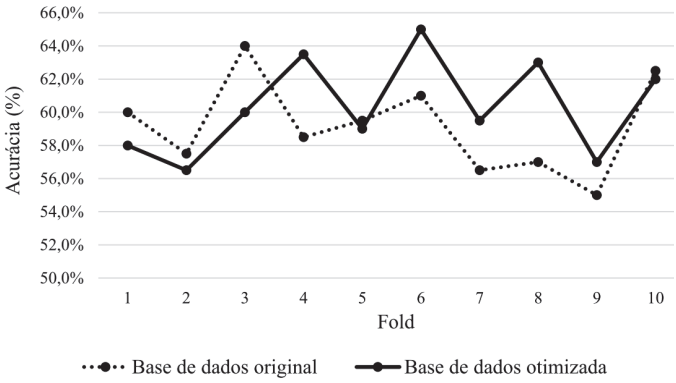

Fig. 2. Comparação entre a acurácia obtida utilizando a base de dados original e base de dados otimizada.

utilizando a base de dados otimizada não apresentou valores melhores se comparados aos da base de dados original, apesar de serem próximos. Realizando a análise estatística destes dados através do teste de Wilcoxon [24], ao comparar o T-*value* calculado com o T-*value* da tabela de valores críticos, e considerando um nível de confiança $\alpha = 0.05$ e $N = 10$ (*N* é a quantidade de *folds*), verificamos que a hipótese nula foi aceita, o que demonstra que os valores observados não apresentaram variação significante.

Por fim, utilizando a base de dados otimizada, o classificador obteve um tempo médio de execução de 2 segundos e 53 milissegundos, o que representa uma redução de 69,4% com relação ao valor obtido utilizando a base de dados completa, a saber, 8 segundos e 27 milissegundos. A avaliação destes dados também foi realizada por meio do teste de Wilcoxon [24]. Ao comparar o T-*value* calculado com o T-*value* da tabela de valores críticos, e considerando um nível de confiança $\alpha = 0.05$ e $N = 10$, verificamos que a hipótese nula não foi aceita, o que demonstra que o tempo médio de execução apresentou uma redução significativa.

Estes resultados mostram que o classificador pôde ser executado em menos tempo, mantendo seu poder de classificação. Além disso, não houve alteração no desvio padrão da acurácia obtida nos *folds* (Tabela IV), comprovando, portanto, a estabilidade do classificador.

VIII. CONCLUSÃO

Este trabalho teve como objetivo avaliar o desempenho do algoritmo evolutivo multiobjetivo NSGA-II em realizar seleção de instâncias para construção de uma base de regras reduzida para um SCBRF, considerando o problema de Classificação de Opiniões. Os experimentos foram realizados em duas

etapas. A primeira considerou a geração da base de regras do SCBRF a partir da base de dados completa, avaliando o tempo de execução e acurácia do classificador. A segunda etapa considerou a geração da base de regras após a redução da base de dados pelo AGMO, avaliando a redução na quantidade de regras, o tempo de execução e a acurácia do classificador fuzzy.

Com base nos resultados apresentados foi possível verificar que os valores de acurácia não apresentaram diferenças estatísticas significativas nas duas etapas. Entretanto, a redução do tempo de execução do SCBRF foi significativa, devido ao menor número de regras geradas, em consequência da seleção de instâncias realizada pelo AGMO. Assim, os resultados demonstraram a eficiência do AGMO NSGA-II em obter um subconjunto de dados menor e com potencial de desempenho para classificação similar, e em alguns casos, maior que o original.

Dando continuidade ao trabalho desenvolvido, experimentos serão realizados utilizando outras bases de dados para Classificação de Opinião, como por exemplo, a base *SFU Review Corpus*[1], que contém documentos de *reviews* de 8 categorias diferentes (filmes, livros, carros, etc) e a base *Multi-Domain Sentiment Dataset*[2] disponibilizada por [25], que consiste em *reviews* de 4 tipos de domínios diferentes retirados da Amazon.com (livros, DVDs, eletrônicos e utensílios de cozinha). Contudo, de forma semelhante à base de filmes utilizada neste trabalho, estas bases precisam ser previamente pré-processadas e estruturadas em bases de dados com instâncias e atributos.

REFERÊNCIAS

[1] P. Grandin and J. M. Adán, "Piegas: A systems for sentiment analysis of tweets in portuguese," *IEEE Latin America Transactions*, vol. 14, no. 7, pp. 3467–3473, 2016.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[3] R. P. Bonidia, J. D. Brancher, and R. M. Busto, "Data mining in sports: A systematic review," *IEEE Latin America Transactions*, vol. 16, no. 1, pp. 232–239, 2018.

[4] M. J. Zaki and W. Meira Jr., *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[5] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[6] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on systems, man, and cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[7] N. García-Pedrajas and J. Pérez-Rodríguez, "Multi-selection of instances: A straightforward way to improve evolutionary instance selection," *Applied Soft Computing*, vol. 12, no. 11, pp. 3590–3602, 2012.

[8] A. Fernandez, V. Lopez, M. J. del Jesus, and F. Herrera, "Revisiting evolutionary fuzzy systems: Taxonomy, applications, new trends and challenges," *Knowledge-Based Systems*, vol. 80, pp. 109–121, 2015.

[9] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data mining and knowledge discovery*, vol. 6, no. 4, pp. 393–423, 2002.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[11] M. Antonelli, P. Ducange, and F. Marcelloni, "Genetic training instance selection in multiobjective evolutionary fuzzy systems: A coevolutionary approach.," *IEEE Trans. Fuzzy Systems*, vol. 20, no. 2, pp. 276–290, 2012.

[1]Disponível em: https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html

[2]Disponível em: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[12] C.-F. Tsai, W. Eberle, and C.-Y. Chu, "Genetic algorithms in feature and instance selection," *Knowledge-Based Systems*, vol. 39, pp. 240–247, 2013.

[13] M. Fazzolari, B. Giglio, R. Alcalá, F. Marcelloni, and F. Herrera, "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, pp. 32–41, 2013.

[14] F. Bertoni and M. Pires, "Aplicação de algoritmos evolutivos multiobjetivo na seleção de instâncias," in *Proceedings [of the] XIII Brazilian Symposium on Information Systems SBSI 2017*, pp. 261–268, jun 2017.

[15] V. M. Bastos, "Ambiente de descoberta de conhecimento na web para a língua portuguesa," *Monograph (Doctored), Federal University of Rio de Janeiro, Rio de Janeiro*, 2006.

[16] O. Cordón, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *International Journal of Approximate Reasoning*, vol. 20, no. 1, pp. 21–45, 1999.

[17] D. Kalyanmoy, *Multi objective optimization using evolutionary algorithms*. John Wiley and Sons, 2001.

[18] C. A. C. Coello, "A short tutorial on evolutionary multiobjective optimization," in *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*, EMO '01, (London, UK, UK), pp. 21–40, Springer-Verlag, 2001.

[19] M. Cardoso, A. Loula, and M. G. Pires, "Automated fuzzy system based on feature extraction and selection for opinion classification across different domains," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 24, no. Suppl. 2, pp. 93–122, 2016.

[20] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[21] B. Ohana and B. Tierney, "Sentiment classification of reviews using sentiwordnet," in *The 9th. IT and T Conference*, jan 2009.

[22] J. R. Cano, F. Herrera, and M. Lozano, "Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study," *IEEE transactions on evolutionary computation*, vol. 7, no. 6, pp. 561–575, 2003.

[23] B. L. Miller, D. E. Goldberg, *et al.*, "Genetic algorithms, tournament selection, and the effects of noise," *Complex systems*, vol. 9, no. 3, pp. 193–212, 1995.

[24] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[25] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

Tayane Leite Cerqueira possui graduação em Bacharelado em Engenharia da Computação pela Universidade Estadual de Feira de Santana (2018). É engenheira de software e gerente de equipe na empresa OpenSEV, em Feira de Santana, Bahia, onde trabalha com programação *full-stack* para sistemas web.

Fabiana Cristina Bertoni possui graduação em Bacharelado em Ciência da Computação pela Universidade Federal de Mato Grosso (2002). Obteve o título de Mestre em Ciência da Computação pela Universidade Federal de São Carlos em 2005, e o título de Doutora em Engenharia Elétrica pela Escola de Engenharia da USP de São Carlos, em 2007. É professora Titular da Universidade Estadual de Feira de Santana, Bahia, Brasil. Tem experiência na área de Sistemas Inteligentes, atuando principalmente nos seguintes temas: Redes Neurais Artificiais, Algoritmos Genéticos, Sistemas Nebulosos e Colônia de Formigas.

Matheus Giovanni Pires possui graduação em Ciência da Computação pela Universidade Paulista (2001), Mestrado em Ciência da Computação pela Universidade Federal de São Carlos (2004) e Doutorado em Engenharia Elétrica pela Escola de Engenharia de São Carlos da Universidade de São Paulo (2009). Atualmente é Professor Titular da Universidade Estadual de Feira de Santana (UEFS). Tem experiência na área de Ciência da Computação, com ênfase em Sistemas Inteligentes, atuando principalmente nos seguintes temas: Aprendizado Genético de Sistemas Fuzzy, Análise de Sentimentos e Mineração de dados.