Seleção de Atributos Agressiva e Efetiva usando Programação Genética*

Felipe Viegas[1], **Isac Sandin**[1], **Thiago Salles**[2] **e Leonardo Rocha**[1]

[1] DCOMP/UFSJ - São João del-Rei, MG , Brasil

[2]DCC/UFMG/- Belo Horizonte, MG , Brasil

(fviegas,isacsandin,lcrocha)@ufsj.edu.br, tsalles@dcc.ufmg.br

Abstract. *A major challenge in automatic classification is to deal with scenarios of high dimensionality. Several feature selection (FS) strategies have been proposed for dimensionality reduction. However, they potentially perform poorly in face of unbalanced data. In this work, we propose a FS strategy based on Genetic Programming in order to overcome this issue. The proposed strategy aims at combining the feature sets selected by distinct FS metrics in order to obtain a more effective set of most discriminative features. We show that our proposal is able to dramatically reduce the data dimensionality, while achieving a more accurate classification.*

Resumo. *Um dos grandes desafios em classificação automática é lidar com cenários de alta dimensionalidade. Várias estratégias de redução de dimensionalidade, incluindo métricas populares de seleção de atributos, já foram propostas, entretanto sem se mostrar adequadas para casos em que os dados são muito desbalanceados. Assim, apresentamos nesse trabalho uma proposta baseada em Programação Genética que visa combinar os resultados de diferentes métricas de seleção de atributos em novos conjuntos, obtendo uma estimativa menos tendenciosa do poder discriminativo de cada atributo. Por meio dessa estimativa conseguimos reduzir a dimensionalidade de forma mais adequada, obtendo resultados de classificação mais precisos.*

1. Introdução

A classificação automática é uma estratégia de aprendizado supervisionado que, dado um conjunto de exemplos previamente rotulados, cria um modelo de aprendizado para classificar novos exemplos (conjunto de teste). Mais formalmente, dado um conjunto de treinamento $\mathbb{D}_{treino} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, onde x_i representa um exemplo e y_i a sua classe, o objetivo é aprender um modelo de classificação que prevê o rótulo de exemplos em $\mathbb{D}_{teste} = \{(x'_1, ?), \cdots, (x'_k, ?)\}$. Existe uma grande variedade de algoritmos propostos para a classificação automática, e vários desafios continuam a receber uma atenção significativa da comunidade científica, tais como lidar com alta dimensionalidade (muitos atributos definindo os objetos) e dados assimétricos (desbalanceamento entre classes).

Particularmente, a aprendizagem baseada em dados com alta dimensionalidade (também conhecido como o problema $p \gg N$, onde p denota a dimensão do espaço de entrada e N denota o número de exemplos de treinamento) é, sem dúvida, um

*Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, e INWEB.

dos maiores desafios na pesquisa de aprendizagem de máquina. À medida em que a dimensionalidade p aumenta, o número de instâncias rotuladas necessárias para produzir um modelo preciso também aumenta, mas de uma forma exponencial [Hastie et al. 2001]. Em essência, quando é feita a aprendizagem com dados de alta dimensionalidade, a identificação de padrões e/ou regiões densas no espaço de entrada pode se tornar uma tarefa muito complexa, motivando o uso de técnicas de redução da dimensionalidade antes dos algoritmos de aprendizagem de máquina. Várias estratégias de redução de dimensionalidade já foram propostas, como a filtragem baseada no poder discriminativo dos atributos—conhecida como Seleção de Atributos ou, em inglês, *Feature Selection* (FS). Nesse caso, o poder discriminativo dos atributos pode ser estimado por uma gama de métricas. Tais métricas estimam uma pontuação para cada atributo, a fim de medir a sua importância na discriminação de classes [Forman 2003]. Assim, uma FS eficaz contribui não apenas para a eficiência da aprendizagem, devido ao menor uso de memória e exigências de processamento, mas também para a eficácia da aprendizagem, uma vez que características menos informativas ou ruídos são filtrados.

Apesar dos esforços acima mencionados, estas métricas não têm se mostrado adequadas para casos em que os dados são muito desbalanceados, uma situação comum no mundo real. Assim, neste trabalho, propomos uma estratégia mais eficaz de FS capaz de lidar com cenários onde os dados sejam desbalanceados. Nossa estratégia se baseia no uso de Programação Genética, em inglês, *Genetic Programming* (GP), que se propõe a procurar, por todo o espaço de combinações possíveis de um conjunto fixo de métricas de FS básicas (e.g., Ganho de Informação, χ^2, Odds Ratio e Coeficiente de Correlação), a combinação que obtém uma estimativa menos tendenciosa do poder discriminativo de cada um dos atributos. O objetivo é tirar proveito das características de cada uma das métricas utilizadas. Avaliamos nossa estratégia em uma coleção de dados extremamente desbalanceada (k8 [Danziger et al. 2009]), relacionada a exemplos mutantes da proteína $p53$, supressora do câncer, classificados em "ativos" e "inativos" quanto à sua função. Nosso resultados experimentais mostram que nossa solução não só aumenta a eficiência dos algoritmos de aprendizagem (com uma redução de cerca de 98% no tamanho do coleção), mas também aumenta significativamente a sua precisão no processo de classificação de amostras de $p53$ (com ganhos de 34% em termos de MacroF_1).

Todas as implementações e execuções de experimentos foram realizadas pelo aluno Felipe Viegas, sob a orientação do professor Leonardo Rocha. As análises de resultados foram feitas em conjunto, aluno e professor. Contou-se também com o auxílio do aluno de doutorado Thiago Salles, do Programa de Pós-Graduação do DCC/UFMG, na compreensão das métricas de FS e com o auxílio do aluno Isac Sandin na configuração do ambiente de Programação Genética. Por fim, é importante mencionar que este trabalho está inserido em um projeto de pesquisa maior que visa combinar várias fontes de dados para se gerenciar com mais eficiência a informação.

2. Trabalhos Relacionados

Nesta seção fazemos uma breve revisão de trabalhos que tiveram como objetivo analisar as métricas de FS para utilização na filtragem de atributos, principalmente em dados assimétricos. Uma métrica de FS é utilizada para designar uma pontuação para cada atributo, a fim de avaliar a sua importância na tarefa de aprendizagem.

Em [Zheng et al. 2004], os autores analisaram certas tendências associadas com as métricas Ganho de Informação (*Information Gain*, ou IG), χ^2, *Odds Ratio* (OR) e Coeficiente de Correlação (*Correlation Coefficient*, ou CC). Segundo os autores, as métricas de FS podem ser agrupadas em 2 conjuntos, unilateral e bilateral. O primeiro conjunto corresponde às métricas que selecionam os atributos mais positivamente significativos para inferir se um determinado objeto pertence a uma classe (isto é, atributos positivos), enquanto o segundo conjunto corresponde às métricas que tanto selecionam os atributos mais significativos para inferir que um objeto não pertence a uma classe (isto é, atributos negativos) quanto os atributos positivos. No conjunto das métricas bilaterais temos a IG e a χ^2 enquanto que no conjunto das métricas unilaterais temos a CC e OR.

Segundo [Forman 2003], os atributos negativos são realmente importantes para a aprendizagem de modelos mais precisos, limitando a aplicabilidade de métricas unilaterais, que os filtram completamente. As métricas bilaterais, quando aplicadas em conjuntos de dados balanceados, selecionam ambos os atributos positivos e negativos, com uma proporção similar à distribuição observada na coleção. No entanto, quando se lida com dados assimétricos, isto não se mantém: métricas bilaterais apresentam uma tendência para os atributos positivos. Assim, motivados pelo desafio de determinar um estimador imparcial para pontuar atributos, que funcione bem mesmo em ambientes altamente assimétricos, objetivamos nesse trabalho, a partir de um conjunto de métricas básicas de FS, empregar GP a fim de encontrar uma maneira de combinar tais métricas de forma mais eficaz. Na próxima seção, fornecemos os detalhes de nossa proposta.

3. Solução Proposta

Nesta seção descrevemos nossa proposta para modelar o problema de FS por meio de Programação Genética (GP). Primeiro, apresentamos a motivação para a utilização de tal abordagem evolucionária para o problema de FS. Em seguida, descrevemos, em termos gerais, o funcionamento de um algoritmo GP. Por fim, detalhamos a nossa estratégia de modelagem para resolver o problema alvo de FS.

Conforme destacado nas seções anteriores, há muitas métricas de FS propostas na literatura. Cada uma dessas métricas pode selecionar diferentes conjuntos de atributos, uma vez que exploram diferentes critérios ao estimar o poder discriminativo dos atributos. Ressaltamos que esses diferentes conjuntos selecionados por tais métricas podem conter bons atributos discriminatórios, bem como atributos não tão relevantes. Assim, nossa expectativa é que, ao combinar as métricas básicas, podemos encontrar um estimador imparcial que alcança uma melhor proporção de atributos selecionados na coleção de dados analisada. Claramente, o espaço de busca sobre todas as possíveis combinações de métricas básicas de FS é extensa, e uma pesquisa por força bruta não é uma escolha sábia. Assim, nossa proposta é empregar o GP para orientar a busca de forma mais eficiente.

GP é um algoritmo evolutivo extensivamente usado (usualmente, com sucesso) para resolver problemas complexos de otimização e de aprendizagem, onde o espaço de busca para uma solução ótima é proibitivamente grande. Ele funciona imitando o processo de evolução de uma população de indivíduos, seguindo o princípio de Darwin da sobrevivência dos mais aptos. Cada indivíduo é normalmente representado por uma árvore, composta por nós terminais (nós folha) e não-terminais (funções, por exemplo, aritméticas ou operadores lógicos). A evolução da população é impulsionada pela

geração e combinação dos seus indivíduos, de acordo com a função saúde (*fitness*), que define a aptidão de um indivíduo. As seguintes etapas resumem o processo de evolução de uma população em GP:

1. Geração aleatória de indivíduos (população inicial), utilizando as funções e os terminais disponíveis. A estratégia mais utilizada é a chamada geração *ramped half-and-half* [Koza 1992].
2. A geração iterativa de uma nova população é realizada através da seguintes subetapas, até que um critério de parada seja atendido:
    (a) Avaliação da *fitness* de cada indivíduo, de acordo com o problema em questão (isto é, a qualidade da solução representada por eles).
    (b) Seleção probabilística (baseada na *fitness*) de um ou dois indivíduos da população para participar das operações genéticas detalhadas em (c). A estratégia mais comumente utilizada é a seleção por torneio, que também será aplicada nesse trabalho.
    (c) Criação de um novo indivíduo usando qualquer um dos seguintes operadores genéticos:
        i. Reprodução: Os indivíduos selecionados são copiados para a nova população.
        ii. Mutação: Um novo indivíduo é criado e adicionado à nova população após uma alteração aleatória em algum nó da árvore.
        iii. *Crossover*: Um novo indivíduo é criado e adicionado à nova população após a recombinação de partes de dois indivíduos escolhidos aleatoriamente.
3. Em cada população, o melhor indivíduo criado é guardado como o resultado da iteração. Uma vez satisfeito o critério de parada, o melhor indivíduo produzido é designado como a solução (talvez aproximada) para o problema.

Lembremos que o problema que visamos resolver é o de determinar os subconjuntos de atributos que melhor expressam as características das classes, em última análise, produzindo uma representação mais compacta dos exemplos de entrada sem diminuir a eficácia de algoritmos de aprendizagem. Para adaptar o arcabouço de GP a esse problema, modelamos cada indivíduo como uma possível combinação entre um conjunto de métricas "básicas" de FS, deixando que o GP busque por indivíduos que produzam a combinação de métricas mais eficaz.

Mais especificamente, considere uma representação de um indivíduo em uma árvore adotada pelo arcabouço de GP. Cada nó terminal consiste de uma métrica de FS ("base"), que retorna um conjunto de atributos considerados altamente discriminativos por tal métrica (ou seja, uma função $f : \mathbb{D}_{treino} \mapsto \mathbb{S}$, onde a entrada $\mathbb{D}_{treino}$ é o conjunto de treinamento e a saída $\mathbb{S}$ o conjunto de atributos mais discriminativos, de acordo com a métrica associada a f). Duas métricas de FS f_i e f_j (nós irmãos na árvore) são combinadas de acordo com uma operação de conjunto especificado pelo nó pai (não-terminal) das mesmas. As operações de conjunto podem ser união ($\cup$), interseção ($\cap$), diferença ($\setminus$), e assim por diante.

Durante o processo de evolução do GP, a qualidade de cada indivíduo é avaliada de acordo com uma função de *fitness* pré-definida. Em nossa proposta, a função de *fitness* é definida com base na qualidade da classificação obtida após a filtragem da coleção de

acordo com os atributos selecionados por um indivíduo (ou seja, obtidos após a aplicação de todos os operadores do indivíduo). Dessa forma, temos que o arcabouço de GP tenta maximizar a função de *fitness*, gerando indivíduos tais que o subconjunto associado aos atributos levam a uma classificação mais precisa. Em seguida, apresentaremos a avaliação experimental da nossa abordagem.

4. Avaliação Experimental

Nesta seção, descrevemos a configuração experimental adotada para avaliar a nossa abordagem. Começamos detalhando as bibliotecas usadas, métricas de FS exploradas, e as métricas utilizadas para avaliar a qualidade da classificação.

4.1. Coleção de Dados e Métricas de Seleção de Atributos

Adotamos em nossos experimentos a coleção $k8$ [Danziger et al. 2009], um conjunto de dados extremamente desbalanceado relacionado a exemplos mutantes da proteína $p53$, supressora do câncer, classificados em "ativos" e "inativos" quanto à sua função. Essa coleção é caracterizada por 5.408 atributos compostos por: características $2D$ provenientes de propriedades da superfície da proteína (4.826 atributos); e características $3D$ que representam os deslocamentos espaciais de resíduos relacionados ao tipo original da proteína (582 atributos). Essa coleção é composta por 16.715 exemplos, classificados como ativo (143 instâncias) e inativo (16.572 instâncias). Note que, quando analisamos essa coleção de dados, o "objetivo final" (em uma configuração de classificação) é discriminar a menor classe (ou seja, a classe "ativa").

As métricas de Seleção de Atributos (*Feature Selection*) utilizadas foram: Ganho de Informação (GI) [Sebastiani 2002], χ^2 [Forman 2003], *Odds Ratio* (OR) [Mladenic 1998] e Coeficiente de Correlação (CC) [Sebastiani 2002]. A seguir apresentamos uma breve descrição de cada uma das métricas. A seguinte convenção de notação será utilizada: denotamos *t* e $\overline{t}$ como a presença ou ausência de um atributo, respectivamente; as classes positiva ("ativa") e negativa ("inativa") são denotadas por *c* e $\overline{c}$. A probabilidade de um atributo *t* ocorrer em uma classe *c* é representado por $P(t|c)$. E as probabilidades $P(c)$ e $P(t)$ indicam a ocorrência de uma classe e de um atributo, respectivamente. Finalmente, N denota o número de exemplos de entrada. Uma breve descrição de cada métrica é apresentada abaixo:

1) Ganho de Informação: quantifica o quanto de informação obtemos sobre uma classe ao saber que um determinado atributo existe ou não em um dado. É dada por $GI(t, c_i) = \sum_{c \in \{c_i, \overline{c}_i\}} \sum_{t' \in \{t, \overline{t}\}} P(t', c) log \frac{P(t', c)}{P(t').P(c)}$.

2) χ^2: é uma métrica geralmente usada em análise estatística para testar se dois eventos são independentes. No contexto de FS, ela é usada para medir a associação entre atributos e classes. Tal métrica é definida por $X^2(t, c_i) = \frac{N[P(t,c_i)P(\overline{t},\overline{c}_i) - P(t,\overline{c}_i)P(\overline{t},c_i)]^2}{P(t)P(\overline{t})P(c_i)P(\overline{c}_i)}$.

3) Coeficiente de Correlação: é usada para estimar a correlação entre classes, e intercorrelação entre atributos. É uma variação da métrica χ^2, onde $CC^2 = \chi^2$. Essa métrica pode ser vista como uma versão "unilateral" de χ^2, sendo definida por $CC(t, c_i) = \frac{\sqrt{N}[P(t,c_i)P(\overline{t},\overline{c}_i) - P(t,\overline{c}_i)P(\overline{t},c_i)]}{\sqrt{P(t)P(\overline{t})P(c)P(\overline{c}_i)}}$.

4) Odds Ratio: mede as chances de um atributo ocorrer em uma classe positiva normalizada pela chances de ocorrer na classe negativa. A ideia básica é que a distribuição

de atributos de um documento relevante é diferente da distribuição de atributos de documentos não relevantes. Ela é dada por $OR(t, c_i) = log\frac{P(t|c_i)[1-P(t|\overline{c}_i)]}{[1-P(t|c_i)]P(t|\overline{c}_i)}$.

4.2. Algoritmos de Classificação e Métricas de Avaliação

Como estamos lidando com atributos contínuos, o algoritmo de classificação automática utilizado foi o Naïve Bayes Gaussiano [Hastie et al. 2001]. Este classificador foi escolhido em função de sua boa eficiência, uma vez que, durante as iterações do GP, várias tarefas de classificação são executadas e avaliadas. A fim de avaliar a eficácia da nossa abordagem, utilizamos as medidas comumente adotadas pela comunidade de Mineração de Dados e Recuperação de Informação, ou seja, microF_1 (Mic.F_1) e macroF_1 (Mac.F_1). A Mic.F_1 mede a eficácia global em termos de todas as decisões feitas pelo classificador (isto é, o inverso da taxa de erro). A Mac.F_1, por outro lado, mede a eficácia de classificação em relação a cada classe de forma independente, calculada pela média harmônica entre precisão e o *recall* obtidos para cada classe [Lewis 1995].

Como mencionado na seção 3, a função de *fitness* é dependente do problema sendo solucionado, e deve refletir o objetivo final de otimização/aprendizagem. Neste trabalho, nosso objetivo principal é fornecer uma classificação de alta qualidade. Além disso, como nossa meta é lidar com dados altamente desbalanceados, usamos a métrica Mac.F_1 para compor a função de *fitness*. De fato, a métrica Mac.F_1 capta melhor a eficácia de classificação para cada classe individualmente (ao contrário da Mic.F_1, que avalia a eficácia global da classificação). Isso se torna importante quando se lida com dados desbalanceados, uma vez que a eficácia em discriminar as classes minoritárias também é levado em conta.

4.3. Biblioteca de Programação Genética

Usamos a biblioteca gpc++ $v0.5.2$ [Weinbrenner 1997], uma biblioteca de GP eficiente para implementar nossa abordagem. Como é uma biblioteca genérica, somente a implementação das estruturas estreitamente relacionadas com o problema foram necessárias, tais como nós terminais, nós de função e a função de *fitness* (em nosso caso, uma única métrica para avaliar a eficácia da classificação, ou seja, a métrica Mac.F_1).

A fim de encontrar os parâmetros utilizados em nossos experimentos, foi realizado um estudo piloto, inspirado pelos resultados relatados pelo autor da biblioteca [Weinbrenner 1997]. Notamos que tal estudo não se encaixa em todos os casos (já que é dependente tanto do problema quanto dos dados), mas, como argumenta o autor, é um bom ponto de partida em direção a uma configuração ideal. Primeiro, foram amostrados 10% do conjunto de dados de maneira aleatória, mantendo a distribuição original das classes. Aplicamos as métricas de FS em tais dados a fim de encontrar o poder discriminativo de cada atributo. Usando os top 5% atributos mais discriminativos, para cada métrica isolada, foi aplicado o procedimento do GP, variando os parâmetros a fim de encontrar os valores que maximizam a *fitness* dos indivíduos (ou seja, aqueles que fornecem a classificação mais precisa). Variamos cada parâmetro do GP de acordo com a seguinte estratégia: O tamanho da população foi ajustado a partir de 50 até 100, com intervalos de 10. O número de gerações foi variado de 20 a 50, com intervalos de 10. A probabilidade de cruzamento foi variada de 90% a 100%, com intervalos de 1%, enquanto tanto as probabilidades de mutação de troca e mutação de redução variaram

de 20% a 100%, com intervalos de 10%. Finalmente, o tamanho do torneio utilizado para a seleção de indivíduos foi escolhido entre 5 e 10, com intervalos de 1. O ajuste de parâmetros baseou-se em um experimento simples em que, ao variar um deles, os demais são mantidos fixos. A configuração que produziu os melhores resultados encontram-se na Tabela 1, sendo esses os valores utilizados em nossos experimentos. Apesar dos bons resultados obtidos na seção seguinte, acreditamos que um ajuste fino destes parâmetros pode levar a resultados ainda melhores. Deixamos esse estudo como trabalho futuro.

Parâmetros	Valores
Tamanho da População	100
Número de Gerações	30
Tipo de Criação	Ramped Half-and-Half
Probabilidade de *Crossover*	98
Tipo de Seleção	Torneio
Tamanho do Torneio	7
Probabilidade de Mutação com *Swap*	50
Probabilidade de Mutação com *Shrinking*	50
Elitismo	Verdadeiro

Tabela 1. Parâmetros do GP

4.4. Discussões e Resultados

Nesta seção, descrevemos os experimentos realizados para avaliar a eficácia da nossa estratégia de combinar métricas de FS baseado em GP. Em seguida, discutimos os resultados obtidos, de acordo com a qualidade da classificação das amostras de $p53$, obtidos após a filtragem dos atributos de cada métrica explorada, bem como aqueles selecionados pela nossa estratégia. A figura 1 ilustra os experimentos realizados, descritos a seguir.

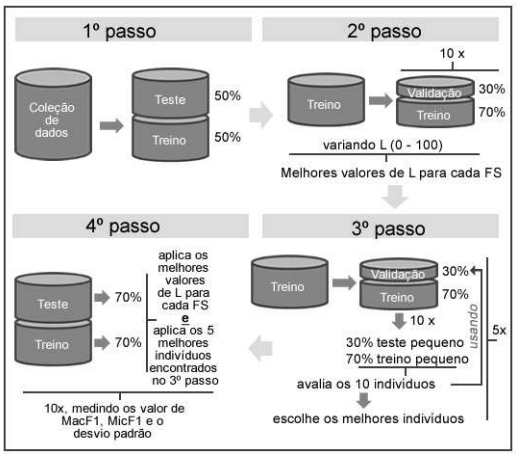
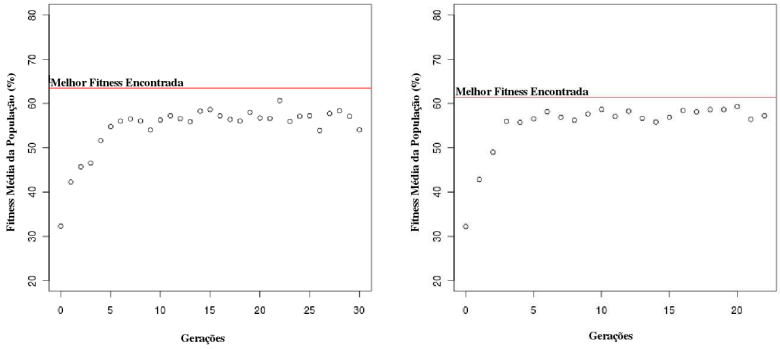


Figura 1. Arquitetura Experimental em Detalhe.

Primeiro, dividimos a coleção em duas partes iguais: Teste e Treino. A primeira partição corresponde ao conjunto de exemplos utilizados na avaliação das métricas e a segunda partição corresponde ao conjunto de exemplos utilizados no processo de calibração das métricas de FS, bem como da nossa estratégia. Este passo de particionamento foi realizado de forma aleatória, mantendo a distribuição original das classes em cada metade.

A segunda etapa consiste em, para cada métrica individual de FS, encontrar o subconjunto de atributos com maior poder discriminativo, via validação cruzada. Cada

uma destas métricas, quando aplicados a um conjunto de dados, pontua cada atributo de acordo com seu poder discriminativo. O objetivo é, portanto, encontrar os $L\%$ atributos com maior poder discriminativo (ou seja, maior pontuação) que melhoram a eficácia da classificação. Mais especificamente, considerando o conjunto de Treino, aplicamos um processo de *hold-out* $70/30$, com 10 repetições, variando L de 5% a 100%, com intervalo de 5%. Os melhores resultados foram obtidos quando L foi ajustada para 10%, 15%, 10% e 10% para as métricas OR, GI, χ^2 e CC, respectivamente.



(a) Indivíduo 1. (b) Indivíduo 2.

Figura 2. Avaliação da *Fitness* média da População.

Tendo determinado os valores de L para cada métrica, o terceiro passo é encontrar os melhores indivíduos. Para isso, utilizando os valores de L determinados, geramos o conjunto de atributos correspondente de cada métrica, os quais serão utilizados no processo de evolução no GP. Dividimos o Treino em dois subconjuntos: Treino e Validação, adotando um processo de *hold-out* com 70% e 30% das amostras, respectivamente. Repetimos esse processo 10 vezes. Para cada repetição, aplicamos o arcabouço de GP e os melhores indivíduos foram armazenados. Em seguida, os melhores indivíduos encontrados em todas as repetições foram avaliados considerando o conjunto de Validação, a fim de selecionar o melhor (ou seja, aquele com maior capacidade de generalização). Este passo foi repetido 5 vezes, resultando em 5 melhores indivíduos. Na figura 2, mostramos a evolução da *fitness* média da população durante o processo evolutivo do GP. Devido às limitações de espaço, relatamos tal análise para 2 dos 5 indivíduos, e destacamos que o comportamento observado foi similar para todos os indivíduos. Como podemos observar, com o processo de evolução contínua, a qualidade da população (em termos de média de *fitness* da população) aumenta, indicando que o GP é, de fato, capaz de seguir um caminho no espaço de busca para um máximo (que pode ser local ou global). Como veremos a seguir, mesmo que em um máximo local, os resultados alcançados em termos de eficácia de classificação são melhores do que usando as métricas tradicionais isoladamente. Os 5 melhores indivíduos encontrados são relatados nas duas primeiras colunas da Tabela 3.

O último passo é finalmente avaliar, considerando o conjunto de Teste, a eficácia de classificação após a filtragem dos $L\%$ atributos mais discriminativos encontrados para cada métrica de FS, bem como os 5 indivíduos encontrados no terceiro passo. Para fazer esta avaliação, foram escolhidas amostras aleatórias do conjunto de Teste (70%) e do conjunto de Treino (70%). Este processo foi repetido 10 vezes. Os valores de Mac.F_1

e MicF_1 obtidos, bem como os desvios padrão, são relatados nas Tabelas 2 e 3 para as métricas individuais e para aquelas obtidas via GP, respectivamente.

	L(%)	Mac.F_1(%)	Mic.F_1(%)
NB+CC	10	49.00 ± 0.63	82.52 ± 0.82
NB+χ^2	10	39.79 ± 1.12	61.32 ± 2.25
NB+GI	15	44.99 ± 0.87	73.58 ± 1.68
NB+OR	10	49.66 ± 0.78	83.92 ± 1.16
NB	100	35.66 ± 0.58	52.20 ± 1.56

Tabela 2. Resultados do NB Combinado com as Métrica de FS

Indivíduo	Expressão	Atributos	Mac.F_1(%)	Mic.F_1(%)
1	$\chi^2 \setminus CC$	9	66.65 ± 1.65	98.59 ± 0.12
2	$OR \setminus GI$	6	60.43 ± 0.90	97.59 ± 0.09
3	$OR \setminus GI$	6	56.24 ± 0.70	97.83 ± 0.31
4	$CC \setminus GI$	5	58.37 ± 0.93	98.01 ± 0.14
5	$CC \setminus \chi^2$	9	55.49 ± 0.75	97.82 ± 0.12

Tabela 3. Os Cinco Melhores Indivíduos Encontrados e seus Resultados.

Comparando as tabelas 2 e 3, podemos observar que nossa estratégia de FS baseada em GP proporcionou melhorias significativas sobre as métricas de FS tradicionais. Outro ponto interessante a se destacar é a redução agressiva do número de atributos obtida por cada indivíduo, sem comprometer a eficácia da classificação. Comparando a melhor FS individualmente com a nossa abordagem baseada no GP, reduzimos os atributos de 504 ($L = 10\%$) para apenas 9 atributos. Mesmo selecionando menos atributos, foi possível melhorar a Mac.F_1 de $49,66\%$ para $66,65\%$, e a Mic.F_1 de 83.92% para 98.59%, o que representa uma melhoria substancial (ganhos 34% em Mac.F_1 e 17% em Mic.F_1). Podemos observar que nossa estratégia pode fazer um trabalho melhor na seleção dos atributos mais importantes, tornando a classificação mais eficaz.

	Mac.F_1(%)	StdDev	Mic.F_1(%)	StdDev
GP	66.65	1.65	98.59	0.12
GI	49.83	0.40	98.37	0.55
OR	50.87	1.64	99.13	0.06
χ^2	50.06	0.52	98.82	0.22
CC	50.15	0.72	99.14	0.05

Tabela 4. Comparação Considerando os 9 Atributos mais Importantes.

Por fim, também comparamos nossa abordagem com as métricas tradicionais de FS, considerando o mesmo número de atributos mais discriminativos encontrados por nossa estratégia (9 atributos), como relatado na Tabela 4. Considerando os 9 atributos melhor pontuados pela métrica OR, foram alcançados Mac.F_1 de $50,87 \pm 1,64$ e Mic.F_1 de $99,13 \pm 0,06$. Ressaltamos aqui que, embora seja observado um ganho substancial em Mic.F_1, tal ganho não é tão expressivo em Mac.F_1, indicando que o modelo obtido tornou-se bastante enviesado (privilegiando a classificação de amostras para a classe majoritária). Por outro lado, com o mesmo número de atributos, a estratégia baseada em GP foi capaz de atingir um valor significativamente maior de Mac.F_1 (com empate estatístico em Mic.F_1), indicando uma maior efetividade na discriminação de amostras da classe minoritária. Esses resultados indicam que o modelo de classificação aprendido, após a filtragem dos atributos, foi mais eficaz para classificar exemplos de teste que pertencem à classe minoritária, explicitando a qualidade do modelo aprendido. Nessa tabela, podemos observar que a abordagem de GP superou todas as métricas tradicionais de FS. Na verdade, o GP foi capaz de produzir um conjunto muito mais eficaz de atributos, corroborando a argumentação feita anteriormente.

5. Conclusões e Trabalhos Futuros

Neste trabalho, propomos uma solução geral para uma estratégia mais eficaz de seleção de atributos que, além de proporcionar uma seleção de atributos altamente eficaz, também é robusta para cenários onde os dados sejam desbalanceados. Nossa solução é baseada no uso de Programação Genética (GP) para aprender uma "combinação" de métricas "básicas" de FS que seja capaz de tirar proveito das características de cada uma delas. Mais especificamente, cada nó terminal consiste de uma métrica "básica" de FS que retorna um subconjunto de atributos considerados mais discriminativos. Duas métricas de FS são combinadas por meio de operadores de conjunto. A função de *fitness* é definida com base na qualidade da classificação obtida após a filtragem da coleção de acordo com os atributos selecionados por um indivíduo. Avaliamos nossa estratégia em uma coleção de dados extremamente desbalanceada relacionada a exemplos mutantes da proteína $p53$, na qual nossa solução não só aumentou a eficiência dos algoritmos de aprendizagem (com uma redução de cerca de 98% no tamanho da coleção), como também aumentou significativamente a sua precisão no processo de classificação de amostras de $p53$ (com ganhos de 34% e 17% em termos de Mac.F_1 e Mic.F_1 respectivamente, quando comparados com a melhor linha de base).

Como trabalho futuro pretendemos aplicar a nossa solução para outros conjuntos de dados, onde os problemas da alta dimensionalidade e distribuições desbalanceadas sejam ainda mais significativos (por exemplo, conjuntos de dados textuais), realizando um estudo mais completo dos parâmetros do GP.

Referências

Danziger, S. A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., and Lathrop, R. H. (2009). Predicting positive p53 cancer rescue regions using most informative positive (mip) active learning. *PLoS Comput Biol*, 5(9):e1000498.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*. Cambridge, MA, USA.

Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. In *Eighteenth Annual, International ACM-SIGIR Conference*, pages 264–254.

Mladenic, D. (1998). *Machine learning on non-homogeneous, distributed text data*. PhD thesis, University of Ljubljana, Faculty of Computer and Information Science.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Weinbrenner, T. (1997). Genetic programming techniques applied to measurement data. Diploma Thesis.

Zheng, Z., Wu, X., and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6:80–89.