Um classificador para seleção de características aplicado a problemas não-linearmente separáveis

Saulo Moraes Villela, Raul Fonseca Neto
Departamento de Ciência da Computação, Universidade Federal de Juiz de Fora, Brasil
Email: {saulo.moraes, raulfonseca.neto}@ufjf.edu.br

***Resumo*—Este artigo apresenta uma abordagem robusta para o problema de seleção de características aplicado a conjuntos de dados não-linearmente separáveis. Neste sentido, foram realizados estudos comparando-se resultados relativos a utilização de classificadores baseados em funções *kernel*, os quais produzem o processo de seleção em um espaço de mais alta dimensão, denominado espaço *kernel*, com classificadores robustos, que utilizam o conceito de margem flexível e permitem uma tolerância a erros de classificação promovendo a seleção de características diretamente no espaço de entrada. A introdução do processo de flexibilização da margem possibilita a correta classificação de dados, que não sejam linearmente separáveis no espaço de entrada, refletindo em uma melhora do poder de generalização. Tal fato pode ser comprovado pela redução de erros nos testes experimentais. Também, a opção pela minimização da norma L_1 do vetor normal ao hiperplano separador, tornou possível a construção de hipóteses com alto grau de esparsidade. De fato, esta forma de otimização, que apresenta um processo de regularização interna, contribui de forma significativa para uma melhor eficiência do processo de seleção de características. Para a seleção dos melhores subconjuntos, os classificadores foram associados a um algoritmo de busca ordenada que utiliza os valores de margem como medida de avaliação dos subconjuntos candidatos. Foram realizados experimentos para a comprovação da proposta apresentada, tendo-se obtido resultados bastante significativos.**

Keywords*—*Seleção de características, Norma L_1, Busca ordenada, Classificadores de larga margem, Margem flexível

I. Introdução

Neste trabalho aborda-se o problema de seleção de características quando aplicado a conjuntos de dados não-linearmente separáveis. Nas tarefas de classificação binária, associadas ao aprendizado supervisionado, os vetores de entrada podem ter uma alta dimensão relacionada à existência de um grande número de variáveis. Entretanto, somente um número reduzido destas variáveis é relevante ou possuem um alto poder de discriminação para separação das duas classes [1], no sentido de produzirem boas hipóteses com alto poder de generalização. Mesmo classificadores de larga margem, como exemplo as Máquinas de Vetores Suportes (*Support Vector Machines* – SVM), não conseguem lidar de forma favorável com este tipo de problema, ou seja, construir boas hipóteses com dados contendo um grande número de variáveis irrelevantes ou até mesmo redundantes. Esta situação se configura mais criticamente em conjuntos de dados não-linearmente separáveis. Neste caso, a utilização de uma função de mapeamento implícita, com a utilização do truque *kernel* [2], expande enormemente a dimensão do problema para um espaço definido como espaço de *kernel*, tornando atrativo o processo de seleção de características. Também, é notório que a realização da seleção de características no espaço *kernel* pode capturar relações de natureza não-linear nos dados que não são observados por técnicas lineares de seleção de características.

De uma forma geral, o processo de seleção de características pode ser executado segundo três abordagens: filtro, embutido (*embedded*) e envoltório (*wrapper*). Os métodos em filtro realizam um ordenamento das variáveis segundo algum tipo de critério, geralmente associado à medidas estatísticas como estatística-t, obtida da distribuição de Student [3], e critérios de Golub [4] e Fisher [5]. Este ordenamento é realizado previamente e não utiliza informações relacionadas ao classificador. Portanto, este processo não depende do tipo de classificador utilizado. O maior ponto negativo desta abordagem está no fato de que, ao se utilizar medidas estatísticas, os valores das variáveis são computados de forma individual, sem levar em consideração a existência de interdependência entre as mesmas.

A estratégia dos métodos embutidos é baseada na utilização do próprio classificador no sentido de escolher os subconjuntos de características mais significativos. Estes métodos possuem um mecanismo interno de regularização que torna possível a escolha das variáveis mais relevantes. Os algoritmos AROM [6], que minimiza uma aproximação da norma zero do vetor normal ao hiperplano separador, e NSC [7], que se baseia na redução de componentes dos centroides das classes, são exemplos importantes desta abordagem.

Em contraste aos métodos em filtro, os métodos *wrapper* utilizam informações fornecidas pelo algoritmo indutor relacionadas a importância ou relevância das variáveis e também do mérito dos subconjuntos. Normalmente estes métodos são implementados junto com uma estratégia heurística de busca para selecionar os melhores subconjuntos. Um importante exemplo deste método é o algoritmo RFE [8], ou algoritmo de eliminação recursiva, que realiza a eliminação recursiva de características a partir de informações providas por um indutor, como por exemplo um classificador SVM.

Como abordagem alternativa, pode-se citar o método de seleção de características que consiste na combinação de um método em filtro, que avalia a qualidade de subconjuntos, com uma busca heurística. Como exemplo desta variante tem-se o algoritmo CFS [9], ou seleção de características baseado em correlação. Este algoritmo utiliza uma heurística baseada em uma medida de correlação entre as variáveis que pertencem a um subconjunto. Este critério de avaliação é associado a uma estratégia construtiva de busca do tipo *best-first* para selecionar o melhor subconjunto.

O desenvolvimento de soluções para o problema de seleção

de características em dados não-linearmente separáveis é conduzido em duas direções principais: na primeira, mais usualmente empregada, a seleção dos subconjuntos é realizada no espaço de entrada, tomando-se o conjunto de características. Entretanto, os estágios de busca, avaliação e classificação são conduzidos no espaço de mais alta dimensão, com a utilização de uma medida ou de um classificador *kernel*. Na segunda, todo processo de seleção de características é realizado no espaço *kernel*, com o emprego de uma função de mapeamento.

No primeiro caso, o método adota uma estratégia híbrida. Isto é justificado pelo fato de que, usualmente, o mapeamento do espaço de entrada para o espaço *kernel* é implícito e, em alguns casos, a dimensão do espaço *kernel* pode ser infinita quando, por exemplo, é utilizado um *kernel* gaussiano. Esta estratégia consiste na geração dos subconjuntos no espaço original utilizando-se de um processo de busca eliminativo ou construtivo tendo como base informações fornecidas por um algoritmo classificador ou por um método em filtro atuando no espaço *kernel*. A avaliação dos subconjuntos também é realizada no espaço *kernel*. Este processo de solução, que utiliza a precisão do classificador como critério de avaliação e que direciona a escolha dos subconjuntos no espaço de entrada, pode ser visto como uma forma indireta de seleção de características no espaço *kernel*. A classificação de novas instâncias com o melhor subconjunto selecionado também é realizada no espaço *kernel*, sempre com o emprego do truque *kernel*.

No segundo caso, o método seleciona diretamente os subconjuntos de características no espaço *kernel*. Para o desenvolvimento desta técnica, torna-se necessário o desenvolvimento de uma função de mapeamento explícita, bem como a geração de uma base de vetores no espaço *kernel* [10]. Desta forma, os vetores selecionados definem um subespaço no espaço *kernel*. Consequentemente, todos os estágios do processo de seleção de características podem ser desenvolvidos neste subespaço utilizando um número limitado de características do mesmo. O maior incentivo para o uso desta abordagem está no fato de que um subconjunto ótimo para o espaço de entrada, escolhido conforme algum critério selecionado, não é necessariamente ótimo para o espaço *kernel* para este mesmo critério. Também, a existência de relações não-lineares entre dados pode ser capturada somente no espaço *kernel*, pelo fato de que não é possível capturar a natureza não-linear destas relações no espaço de entrada do problema. A classificação de novas instâncias com o melhor subconjunto de características selecionado também é realizada no subespaço *kernel*.

O método proposto neste trabalho considera uma simples estratégia que consiste na execução de todo processo de seleção de características no espaço de entrada do problema. De acordo com Vapnik [11], isto se torna possível pela utilização de uma margem flexível no classificador, permitindo a relaxação das restrições de classificação. A fim de tornar mais rápido e eficiente o processo de busca, optou-se pela utilização de um classificador que minimiza a norma L_1 do vetor normal ao hiperplano separador assegurando a geração de hipóteses com maior grau de esparsidade. Esta forma de minimização tem como resultado a maximização da distância L_∞ entre os vetores suportes de classes contrárias no espaço de entrada. Os resultados obtidos foram comparados com a utilização de um classificador SVM que minimiza a norma Euclidiana do vetor normal e que, por consequência, maximiza também a distância Euclidiana entre vetores suportes de classes opostas. Para ambos os métodos foi adotada uma abordagem do tipo envoltória baseada na obtenção de valores de margens fornecidos pelos respectivos classificadores utilizados na avaliação dos subconjuntos candidatos. Os classificadores são acoplados ao algoritmo de Busca Ordenada Admissível (*Admissible Ordered Search* – AOS) [12], o qual executa um processo de eliminação sequencial no espaço original. Este processo de eliminação é baseado na escolha das componentes de menor magnitude do vetor normal relacionado ao hiperplano separador de cada hipótese.

Foram conduzidos experimentos em cinco bases de dados não-linearmente separáveis, comparando-se o processo de seleção de características no espaço de entrada com o uso de uma função *kernel* gaussiana. Resultados experimentais validaram a efetividade da proposta em termos de poder de generalização e eficiência computacional.

As seções subsequentes do artigo foram organizadas como segue: a seção II descreve de forma resumida os conceitos preliminares e a base teórica do classificador robusto de margem flexível. A seção III descreve o algoritmo de busca AOS bem como a realização da seleção de características nos espaços de entrada e *kernel*. Na seção IV apresentam-se os resultados dos experimentos computacionais e, finalmente, na seção V algumas considerações e conclusões sobre o trabalho são reportadas.

II. Classificador robusto de margem flexível

A. Problema de classificação binária

Seja $Z = \{z_i = (x_i, y_i) : i \in \{1, \ldots, m\}\}$ um conjunto de treinamento composto de pontos $x_i \in \mathbb{R}^d$ e rótulos (classes) $y_i \in \{-1, +1\}$. Além disso, sejam Z^+ e Z^- definidos como os conjuntos $\{(x_i, y_i) \in Z : y_i = +1\}$ e $\{(x_i, y_i) \in Z : y_i = -1\}$, respectivamente. Um problema de classificação binária consiste em encontrar um hiperplano, que é dado pelo seu vetor normal $w \in \mathbb{R}^d$ e uma constante $b \in \mathbb{R}$, de tal forma que os pontos em Z^+ e Z^- sejam separados nos dois semi-espaços gerados por ele. Assim, define-se (w, b) tal que:

$$y_i \left(w \cdot x_i + b\right) \geq 0, \text{ para todo } (x_i, y_i) \in Z.$$

É dito que Z aceita uma margem $\gamma \geq 0$ quando existe um hiperplano $\mathcal{H} := \left\{x \in \mathbb{R}^d : w \cdot x + b = 0\right\}$ tal que:

$$y_i \left(w \cdot x_i + b\right) \geq \gamma, \text{ para todo } (x_i, y_i) \in Z.$$

Nesse caso, define-se dois hiperplanos adicionais paralelos à $\mathcal{H}$, dados por $\mathcal{H}^+ := \{x \in \mathbb{R}^d : w \cdot x + (b - \gamma) = 0\}$ e $\mathcal{H}^- := \{x \in \mathbb{R}^d : w \cdot x + (b + \gamma) = 0\}$. A distância entre estes dois hiperplanos paralelos sob uma norma p é dada por [13]:

$$\text{dist}\left(\mathcal{H}^+, \mathcal{H}^-\right) = \frac{-b + \gamma + b + \gamma}{||w||_q} = \frac{2\gamma}{||w||_q},$$

onde p e q são normas conjugadas, tal que $\frac{1}{p} + \frac{1}{q} = 1$. Sendo $\gamma_g := \text{dist}(\mathcal{H}^+, \mathcal{H}^-)/2$, define-se γ_g como a margem geométrica de norma p entre os dois hiperplanos $\mathcal{H}^+$ e $\mathcal{H}^-$. Sendo assim, é dito que Z aceita uma margem geométrica ($\gamma_g \geq 0$) quando existe um hiperplano com (w, b) tal que:

$$y_i \left(w \cdot x_i + b\right) \geq \gamma_g ||w||_q, \text{ para todo } (x_i, y_i) \in Z. \qquad (1)$$

B. Perceptron de Margem Fixa com norma p – FMP$_p$

Dado um valor de margem fixa γ_f e um conjunto de treinamento Z, que aceita γ_f como uma margem geométrica, isto é, satisfaz a equação (1), considera-se o problema de encontrar um hiperplano separador com dados (w, b) tal que:

$$y_i \left(w \cdot x_i + b\right) \geq \gamma_f ||w||_q, \text{ para todo } (x_i, y_i) \in Z. \qquad (2)$$

A formulação desse problema, para a norma euclidiana, foi proposta em [14] e o algoritmo resultante foi denominado de Perceptron de Margem Fixa (FMP). Em [15] é apresentada uma extensão do mesmo para uma norma arbitrária e o algoritmo de Perceptron de Margem Fixa com norma p (*Fixed p-Margin Perceptron* – FMP$_p$). Nesse caso, a função de erro, ou perda, $J : \mathbb{R}^{d+1} \to \mathbb{R}$ é definida como:

$$J\left(w, b\right) := \sum_{(x_i, y_i) \in \mathcal{M}} \gamma_f ||w||_q - y_i \left(w \cdot x_i + b\right),$$

onde $\mathcal{M}$ é o subconjunto de Z que viola a equação (2) para os dados (w, b), assim $\mathcal{M} := \{(x_i, y_i) \in Z : y_i \left(w \cdot x_i + b\right) < \gamma_f ||w||_q\}$.

Seguindo a abordagem *on-line* do gradiente estocástico, o processo de minimização começa com valores iniciais (w^0, b^0), normalmente $(0, 0)$. A cada iteração do algoritmo, um par $z_i = (x_i, y_i)$ é escolhido e verificado contra (w^t, b^t), $t \in \{1, 2, \ldots\}$. Se o par z_i escolhido for um erro, ou seja, se $y_i \left(w^t \cdot x_i + b^t\right) < \gamma_f ||w^t||_q$, então um novo vetor normal w^{t+1} e a constante b^{t+1} são construídos usando o gradiente de J. Tomando as derivadas parciais de J em relação à w_j, $j \in \{1, \ldots, d\}$, e à b quando $1 < q < \infty$, tem-se:

$$\begin{aligned} \frac{\partial J\left(w, b\right)}{\partial w_j} &= \gamma_f \left(\sum_{i=1}^{d} |w_i|^q \right)^{\frac{1}{q} - 1} |w_j|^{q-1} \text{sinal}(w_j) - y_i x_{ij} \\ \frac{\partial J\left(w, b\right)}{\partial b} &= -y_i. \end{aligned}$$

Sendo assim, para $1 < q < \infty$, obtém-se a seguinte regra de correção para o FMP$_p$ quando z_i for um erro:

$$\begin{aligned} w^{t+1} &\leftarrow w^t - \eta \left(\gamma_f ||w^t||_q^{1-q} |w^t|^{q-1} \text{sinal}(w^t) + y_i x_i \right) \\ b^{t+1} &\leftarrow b^t + \eta y_i, \end{aligned}$$

onde $\eta \in (0, 1]$ é a taxa de aprendizado, $|w| := \left(|w_i|, \ldots, |w_d|\right)'$, e $\text{sinal}(w) := \left(\text{sinal}(w_i), \ldots, \text{sinal}(w_d)\right)'$.

Para se ter uma regra de atualização para o caso onde $p = \infty$, pega-se o limite $q \downarrow 1$ na expressão da $\partial J / \partial w_j$. A análise é trivial, derivando a seguinte regra de correção:

$$w^{t+1} \leftarrow w^t - \eta \left(\gamma_f \text{sinal}\left(w^t\right) + y_i x_i \right).$$

C. Margem Flexível com norma p

Em problemas onde o conjunto de treinamento apresenta *outliers* torna-se importante o processo de flexibilização da margem. Também é possível obter hiperplanos separadores com uma maior margem, aumentando o poder de generalização. A flexibilidade da margem também pode ser útil no tratamento de dados não-linearmente separáveis. Isto se torna importante quando se utiliza uma norma arbitrária, visto que, nesse caso, não é possível a utilização da formulação dual e, consequentemente, de funções *kernel*.

Neste sentido, apresenta-se nesta seção uma estratégia no sentido de permitir a flexibilização da margem para o algoritmo FMP$_p$. A ideia consiste em adaptar o esquema de flexibilização introduzido por [16] que considera a penalização quadrática das variáveis de folga associada às restrições de classificação. Este problema, originalmente formulado para norma euclidiana, é descrito da seguinte forma:

$$\begin{gathered} \text{Min } \tfrac{1}{2} w \cdot w^t + C \cdot \textstyle\sum_i \varepsilon_i^2 \\ \text{Sujeito a} \\ y_i \left(w \cdot x_i + b\right) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0 \text{ para todo } i \in \{1, \ldots, m\}, \end{gathered}$$

onde ε é o vetor das variáveis de folga e C um parâmetro de controle.

Tomando-se o gradiente da função de Lagrange com respeito aos parâmetros primais obtém-se a seguinte formulação dual equivalente do problema:

$$\begin{gathered} \text{Max } \textstyle\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle - \sum_i \alpha_i^2 / 2 \cdot C \\ \text{Sujeito a} \\ \textstyle\sum_i \alpha_i y_i = 0, \alpha_i \geq 0 \text{ para todo } i \in \{1, \ldots, m\}. \end{gathered}$$

Fazendo $\lambda = 1 / \left(2 \cdot C\right)$, pode-se redefinir o problema dual como a soma deste valor constante à diagonal da matriz *kernel*. Este valor é inversamente proporcional ao parâmetro de controle C. Em outras palavras, reescreve-se a matriz de produtos internos ou matriz *kernel* na forma:

$$\begin{aligned} K\left(x_i, x_j\right) &= \left\langle x_i, x_j \right\rangle, \text{ para todo } j \neq i \\ K\left(x_i, x_i\right) &= \left\langle x_i, x_i \right\rangle + \lambda, \text{ caso contrário.} \end{aligned}$$

Considerando a representação expandida do vetor normal w, pode-se substituir o mesmo no sistema de restrições da formulação primal do algoritmo FMP$_p$, no sentido de expressar as inequações em função do conjunto de multiplicadores α. Assim, tem-se, para uma norma arbitraria p:

$$y_i \left(\sum_{j=1}^{m} \alpha_j y_j K\left(x_i, x_j\right) + b \right) \geq \gamma_f ||w||_q.$$

Introduzindo as variáveis de folga, tem-se:

$$y_i \left(\sum_{j=1}^{m} \alpha_j y_j K\left(x_i, x_j\right) + b \right) + \alpha_i \cdot \lambda \geq \gamma_f \cdot ||w||_q.$$

Portanto, pode-se estimar o valor apropriado das variáveis de folga em função dos multiplicadores associados a cada restrição e do parâmetro de controle λ. Evidentemente, quanto maior o valor de λ maior será o nível de flexibilização das restrições não satisfeitas ou violadas.

Para a norma euclidiana, sempre que ocorrer um erro associado a amostra x_i, o valor do multiplicador associado é corrigido pela equação:

$$\alpha_i = \alpha_i + \eta \cdot 1.$$

Esta correção, deve ser seguida de um escalonamento no vetor de multiplicadores na forma:

$$\alpha^{t+1} \leftarrow \alpha^t \left(1 - \left(\eta \cdot \gamma_f\right) / ||w||_2\right).$$

De acordo com [14], isto se torna equivalente à correção do vetor normal w na formulação primal do algoritmo:

$$w^{t+1} \leftarrow w^t \left(1 - (\eta \cdot \gamma_f) / ||w||_2\right) + \eta \cdot x_i y_i.$$

Para uma norma arbitrária p adota-se uma estratégia que consiste na realização do escalonamento do vetor α com base na respectiva norma conjugada, ou seja:

$$\alpha^{t+1} \leftarrow \alpha^t \left(1 - (\eta \cdot \gamma_f) / ||w||_q\right).$$

Em uma interpretação Bayesiana do processo de flexibilização [17], [18], onde a matriz de produtos internos é vista como a matriz de covariância dos dados, a adição de um valor constante aos componentes da diagonal principal pode ser interpretada como a soma de uma variância relacionada a existência de ruídos ou *outliers* nos valores das respectivas margens funcionais. Entretanto, deve-se adotar uma medida de precaução na atribuição do valor ao parâmetro de controle λ. A existência de valores muito altos associados aos componentes da diagonal principal em relação aos demais valores da matriz induzem a ocorrência de *overfitting*.

D. Algoritmo de Margem Incremental com norma p – IMA$_p$

Uma nova formulação para o problema de maximização da margem foi proposta por Leite e Fonseca Neto [14] e desenvolvida a partir da constatação de que, na obtenção da máxima margem, os pontos ou vetores suporte de classes contrárias se encontram à mesma distância do hiperplano separador, ou seja, considerando as margens das classes de rótulos positivo e negativo, tem-se $\gamma_+^* = \gamma_-^*$, onde:

$$\begin{aligned} \gamma_+^* &= \text{Min } y_i \left(w \cdot x_i + b\right) / ||w||_2, \text{ para todo } x_i \in Z^+ \\ \gamma_-^* &= \text{Min } y_i \left(w \cdot x_i + b\right) / ||w||_2, \text{ para todo } x_i \in Z^-. \end{aligned}$$

Os autores propõem a solução aproximada do problema de máxima margem, considerando a maximização explícita e direta da margem geométrica. Adaptando-se para uma norma p [15], deve-se resolver o seguinte problema de otimização:

$$\begin{gathered} \text{Max}_w \gamma_g \\ \text{Sujeito a} \\ y_i \left(w \cdot x_i + b\right) \geq \gamma_g ||w||_q. \end{gathered}$$

A técnica de solução desenvolvida consiste em uma estratégia de aprendizado incremental, através da qual são obtidas sucessivas soluções do problema FMP$_p$, para valores crescentes de margem. Esse parâmetro inicia com o valor zero e tem seus valores incrementados de forma consistente até aproximar-se do valor da margem máxima. Ou seja, para um conjunto de valores $\gamma_f \in [0..\gamma^*)$, sendo:

$$\gamma_f^{t+1} > \gamma_f^t, \text{ para } t = 1, \ldots, T-1, \gamma_f^1 = 0, \gamma_f^T \approx \gamma^*,$$

soluciona-se o problema de inequações não lineares:

$$y_i \left(w \cdot x_i + b\right) \geq \gamma_f ||w||_q, i \in \{1, \ldots, m\},$$

sendo cada solução equivalente à solução do problema do Perceptron de Margem Fixa com norma p.

Para a atualização, a cada iteração, do valor da margem fixa, adotam-se duas regras, baseadas em uma estratégia de balanceamento, que garantem um direcionamento para a solução de máxima margem:

Primeira regra: caso a solução do problema forneça as margens, negativa e positiva, diferentes, pode-se dizer que a solução obtida não caracteriza uma solução de máxima margem. Portanto, corrige-se o valor da margem fixa na forma:

$$\gamma_f^{t+1} = \frac{\gamma_+^t + \gamma_-^t}{2},$$

onde γ_+^t e γ_-^t são os valores relacionados, respectivamente, às menores distâncias projetadas dos pontos do conjunto Z^+ e Z^- ao hiperplano separador da t-ésima iteração.

Segunda regra: caso a solução do problema forneça as margens, negativa e positiva, iguais, pode ser que a solução obtida seja uma solução de ótimo local. Portanto, torna-se necessário garantir um acréscimo no valor da nova margem fixa, na forma:

$$\gamma_f^{t+1} = (1 + \Delta) \gamma_f^t,$$

sendo $\Delta \in (0, 1)$ uma constante de incremento.

Entretanto, adotando esse incremento, não se tem mais a garantia de solução do novo problema, já que o novo valor da margem fixa poderá ser igual ou maior que o valor da margem ótima, ou seja, $\gamma_f^{t+1} \geq \gamma^*$. Para a solução desse contratempo, é suficiente a imposição de um número máximo de iterações no número de épocas do algoritmo de treinamento. Caso não haja uma nova solução do problema FMP$_p$, adota-se, como margem, o valor da margem fixa anterior, relacionado à última solução. Para tornar mais eficiente o processo de otimização, como solução inicial de um novo problema FMP$_p$, utiliza-se a solução do problema anterior.

III. Busca ordenada admissível

A. Espaço de estados e busca heurística

O algoritmo AOS se baseia em uma busca ordenada admissível e tem a capacidade de encontrar, em cada dimensão do problema, o subconjunto de características que possui maior margem, considerando um espaço de estados restrito. É um método do tipo *wrapper* baseado em um processo de eliminação de características que utiliza como indutor um algoritmo de larga margem, como o algoritmo IMA$_p$. Em um processo de busca ordenada, a admissibilidade do algoritmo é assegurada se for utilizada uma função de avaliação monótona [19]. Portanto, considerando como objetivo a maximização da margem, utiliza-se como mérito final de cada hipótese o valor de margem obtido por um classificador que tem sua dimensão associada ao respectivo subconjunto de características. A admissibilidade do processo é preservada pelo fato de que os valores de margem são não crescentes na medida em que a dimensão do problema é reduzida.

Seja $\gamma_{g_j}^{d-1}$ o valor real da maior margem associada a uma hipótese sucessora ou filha que elimina a j-ésima variável. Seja γ_g^d o valor real da margem da hipótese pai em um espaço de dimensão d. Então, em um espaço de dimensão d-1, tem-se:

$$\gamma_{g_j}^{d-1} \leq \gamma_g^d, \forall j.$$

A estratégia de controle do algoritmo é implementada através da inserção das hipóteses geradas em uma fila de prioridade ordenada em ordem decrescente pelos respectivos valores de margens.

B. Margem projetada

Para inserir na fila de prioridade uma hipótese com o valor real da margem deve-se resolver um problema de maximização de margem semelhante ao problema SVM para cada hipótese gerada. Entretanto, para atenuar este esforço computacional, utiliza-se como valor inicial de margem para cada hipótese um valor otimista conhecido como margem projetada. Este valor, obtido da projeção da margem da hipótese pai no espaço de dimensão inferior, representa um limite superior em relação ao valor real da margem. Neste sentido, estados com valores de margem projetada podem, caso necessário, serem podados ou eliminados da fila de prioridade sem afetarem a admissibilidade do algoritmo. O vetor relativo a margem geométrica em um espaço $\mathbb{R}^d$ é definido como: $\gamma_g^d \cdot w/||w||_p$.

Seja $\gamma_{p_j}^{d-1}$ a margem projetada em $\mathbb{R}^{d-1}$ de uma hipótese relacionada a eliminação da j-ésima característica. Então, para uma norma euclidiana:

$$\gamma_{p_j}^{d-1} = \frac{\gamma_g^d}{||w||_2} \left(\sum_{k \neq j} w_k^2 \right)^{\frac{1}{2}}.$$

É facilmente observável que se a j-ésima componente do vetor normal w tiver valor 0 então o valor da margem projetada terá valor equivalente a margem da hipótese pai. Também, este cálculo do valor da margem projetada pode ser estendido para o caso geral de uma norma arbitrária adotando-se a fórmula de Minkowski.

Portanto, quando uma hipótese que possui o maior valor de margem na fila de prioridade é selecionada, independentemente de sua dimensão, duas hipóteses podem ocorrer:

Primeira: o valor da margem é um valor projetado. Neste caso, calcula-se o valor da margem real e compara-se com o valor de margem associado a próxima hipótese da fila. Se este valor continuar sendo a melhor opção no sentido de ser o maior, remove-se a hipótese da fila e processa-se a geração das hipóteses filhas. Caso contrário, reinsere-se a hipótese em questão com o valor da margem real na posição apropriada.

Segunda: o valor da margem é um valor real. Neste caso, simplesmente remove-se a hipótese da fila e processa-se a geração das hipóteses filhas.

C. Fator e critérios de ramificação

A estratégia de exploração adotada pelo algoritmo AOS consiste na eliminação dos menores componentes associados ao vetor normal w. É importante ressaltar que, quando a componente do vetor tiver o valor 0, o valor da margem projetada se iguala ao valor da margem da hipótese pai e consequentemente ao valor real da hipótese filha. Neste caso não será preciso resolver o problema de maximização da margem e o valor da margem é inserido como valor real. Esta estratégia do algoritmo AOS é de extrema importância quando se considera a minimização da norma L_1 e a geração de hipóteses esparsas.

Como observado, o critério de ramificação está associado a magnitude dos valores das componentes do vetor w. Entretanto, para o caso não-linear, onde se utiliza uma função *kernel*, o vetor w não pode ser expresso explicitamente no espaço *kernel*. Neste caso, utiliza-se um critério equivalente proposto em [20], relacionado ao cálculo da menor variação observada no valor da função objetiva dual do problema de maximização da margem.

Seja J o valor da função objetiva em função do vetor α de variáveis duais:

$$J = \frac{1}{2} \cdot \alpha^T \cdot H \cdot \alpha - \alpha^T \cdot \mathbf{1},$$

onde H representa a matriz *kernel* em sua forma mais geral. Portanto, se for excluída a k-ésima variável, tem-se:

$$\Delta J(k) = J - J(k) = \frac{1}{2} \cdot \alpha^T \cdot \Delta H_k \cdot \alpha = ||w||_2 - ||w_k||_2.$$

Portanto, a seleção para eliminação da variável que produz a menor variação no valor da função J, ou de forma equivalente na norma do vetor w, corresponde ao critério relacionado a seleção da componente de menor magnitude do vetor w.

Para o cálculo do valor da margem projetada no espaço *kernel*, utiliza-se a seguinte expressão derivada da relação entre as respectivas normas do vetor w expressas em função do vetor α:

$$\gamma_{p_j}^{d-1} = \left(\frac{1}{\alpha^T \cdot H \cdot \alpha} \right) \cdot \gamma_g^d \cdot \alpha^T \cdot H_j \cdot \alpha.$$

IV. Experimentos e resultados

A. Conjuntos de dados

Para a análise dos resultados, foram utilizadas cinco bases não-linearmente separáveis. As bases utilizadas neste trabalho estão contidas na UCI Machine Learning Repository [21] e são mostradas na Tabela I. Para a base Wine, que possui 3 classes, as classes foram divididas nos tipos 1 e 3 contra o tipo 2.

Tabela I. Informações das bases de dados.

Base	Atributos	Amostras		
		Pos.	Neg.	Total
Ionosphere	34	225	126	351
WDBC	30	212	357	569
Bupa	6	145	200	345
Pima	8	268	500	768
Wine	13	107	71	178

B. Comparação nas bases de dados

A fim de testar os algoritmos nos conjuntos de dados e validar os resultados, os conjuntos foram divididos em dois subconjuntos, 2/3 para o processo de seleção de características e 1/3 para a estimativa do erro de predição (teste). Para estimar o erro de validação relacionado aos conjuntos de treinamento foi utilizado o método de validação cruzada *k-fold cross-validation* [22]. A fim de estimar o erro de validação final, foi mantida a percentagem de pontos de cada classe de dados e foi calculado o erro médio de 10 execuções de um 10-*fold cross-validation*. Para comparações mais precisas, foram selecionados, para cada base, sempre os mesmos conjuntos de treinamento e teste e sempre os mesmos 10 subconjuntos para as validações cruzadas, preservando a semente geradora associada ao processo de aleatoriedade.

Foram realizados testes de seleção de características utilizando o AOS com o IMA_∞ como indutor, a fim de escolher

um representante desse método para casa base. Foi utilizado, para o algoritmos AOS, um fator de ramificação igual a 3, como sugerido em [12]. Como critério (dimensão) de parada, foi escolhido o valor em que pelo menos uma das larguras escolhidas para o gaussiano conseguiu atingir, uma vez que todas as execuções utilizando o IMA_∞ chegaram pelo menos nas mesmas dimensões das execuções de um classificador SVM com um *kernel* gaussiano. A Tabela II mostra os resultados com o conjunto original e as soluções do AOS juntamente com o IMA_∞. Os valores de flexibilidade foram escolhidos como as menores potências de 10 que obtiveram soluções.

Tabela II. ESCOLHA DO PARÂMETRO DE FLEXIBILIDADE.

Base	Flex.	Conjunto Original			AOS + IMA_∞		
		F	10-*fold*	Teste	F	10-*fold*	Teste
Ionosphere	1	34	12,18%	13,68%	3	15,33%	17,09%
	10		14,87%	17,09%		**14,92%**	**17,09%**
	100		35,19%	35,90%		26,61%	31,62%
WDBC	100	30	8,55%	5,26%	2	11,70%	9,47%
	1000		7,65%	4,74%		**9,90%**	**6,84%**
	10000		8,21%	5,26%		20,43%	10,00%
Bupa	1	6	36,70%	39,13%	3	40,30%	35,65%
	10		35,87%	38,26%		42,57%	44,35%
	100		32,78%	40,00%		**38,09%**	**39,13%**
Pima	10	8	24,89%	30,47%	3	**25,61%**	**26,17%**
	100		30,65%	28,13%		31,07%	27,73%
	1000		33,05%	36,72%		34,80%	35,16%
Wine	1	13	2,70%	8,47%	2	**10,91%**	**8,47%**
	10		7,73%	6,78%		13,58%	8,47%
	100		15,48%	6,78%		23,90%	18,64%

Após essa escolha, foram realizados testes comparativos entre o representante do IMA_∞ e um classificador SVM com *kernel* gaussiano, com duas larguras (σ = 0,01 e σ = 1). Como classificador SVM foi utilizado o algoritmo de Otimização Mínima Sequencial (*Sequential Minimal Optimization* – SMO) [23]. Os resultados são apresentados na Tabela III.

Tabela III. AOS COM IMA_p VERSUS AOS COM SMO.

Base	F	AOS + IMA_∞		AOS + SMO			
				0,01		1	
		10-*fold*	Teste	10-*fold*	Teste	10-*fold*	Teste
Ionosphere	3	**14,92%**	**17,09%**	—		16,37%	23,08%
WDBC	2	**9,90%**	**6,84%**	31,82%	7,89%	37,42%	37,89%
Bupa	3	38,09%	**39,13%**	**37,13%**	41,74%	41,13%	40,87%
Pima	3	**25,61%**	**26,17%**	35,31%	28,52%	34,88%	36,72%
Wine	2	**10,91%**	**8,47%**	22,36%	18,64%	35,01%	22,03%

V. CONCLUSÃO

O presente trabalho propôs a aplicação de um classificador robusto para seleção de características em problemas não-linearmente separáveis. Este classificador realiza a minimização da norma L_1 do vetor normal ao hiperplano separador e permite a flexibilização das restrições de margem. Foi observado, nos experimentos realizados, resultados bastante satisfatórios. Exceto para a base de dados Bupa, onde o erro de validação do processo de seleção de características no espaço *kernel* foi ligeiramente inferior, em todos os outros resultados a estratégia apresentada resultou em erros de validação e teste inferiores. Como trabalhos futuros, propõe-se a utilização do classificador combinado a um método de seleção embutido.

REFERÊNCIAS

[1] A. Y. Ng, "On feature selection: Learning with exponentially many irrelevant features as training examples," in *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, 1998, pp. 404–412.

[2] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY: Cambridge University Press, 2004.

[3] D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer, *Mathematical Statistics with Applications*, 7th ed. Belmont, CA, USA: Duxbury Advanced Series, 2008.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286(5439), pp. 531–537, 1999.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[6] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[7] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[8] I. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[9] M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, 1999, pp. 235–239.

[10] G. Baudat and F. Anouar, "Feature vector selection and projection using kernels," *Neurocomputing*, vol. 55, no. 1-2, pp. 21–38, 2003.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[12] S. M. Villela, S. C. Leite, and R. Fonseca Neto, "Feature selection from microarray data via an ordered search with projected margin," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 3874–3881.

[13] A. Dax, "The distance between two convex sets," *Linear Algebra and its Applications*, vol. 416, pp. 184–213, 2006.

[14] S. C. Leite and R. Fonseca Neto, "Incremental margin algorithm for large margin classifiers," *Neurocomputing*, vol. 71, pp. 1550–1560, 2008.

[15] S. M. Villela, S. C. Leite, and R. Fonseca Neto, "Algoritmo de margem incremental com norma p para classificadores de larga margem," in *XI CBIC - Congresso Brasileiro de Inteligência Computacional*, Porto de Galinhas, PE, 2013.

[16] B. Schölkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2002.

[17] R. Herbrich, *Learning Kernel Classifiers: Theory and Algorithms*, ser. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press, 2002.

[18] J. Shawe-Taylor and N. Cristianini, *Margin Distribution and Soft Margin*, ser. Neural information processing series. Cambridge, MA: MIT Press, 1999, ch. 2, pp. 5–16.

[19] P. Hart, N. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[21] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[23] J. C. Platt, *Fast training of support vector machines using sequential minimal optimization*. Cambridge, MA: MIT Press, 1999, pp. 185–208.