X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

Seleção de Características utilizando Busca Ordenada e um Classificador de Larga Margem

Saulo Moraes Villela, Adilson Elias Xavier
Programa de Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro
Caixa Postal 68.511, 21941-972 Rio de Janeiro, RJ, BRASIL
{saulomv,adilson}@cos.ufrj.br

Raul Fonseca Neto, Saul de Castro Leite
Departamento de Ciência da Computação, Universidade Federal de Juiz de Fora
36036-330 Juiz de Fora, MG, BRASIL
{raulfonseca.neto,saul.leite}@ufjf.edu.br

Resumo – Máquinas de Vetores Suportes têm sido amplamente utilizadas com bastante eficiência em problemas que possuem uma grande quantidade de variáveis ou características. Neste sentido, torna-se importante o desenvolvimento de novas estratégias para a solução deste tipo de problema que se utilizem de critérios de seleção associados a este tipo de classificador. Neste trabalho propõe-se um novo método para a seleção de subconjuntos de variáveis que utiliza um processo de busca ordenada, também conhecido como *best-first*, para exploração do espaço de possíveis candidatos. O algoritmo, denominado AOS, utiliza como medida de avaliação os valores de margem calculados a partir da utilização de um classificador de larga margem, denominado IMA. Este classificador, de grande eficiência computacional, permite grande flexibilidade e rapidez na obtenção dos valores de margem possibilitando a solução de problemas de tamanho razoável, com centenas de características, sem que ocorra explosão combinatória. O algoritmo foi testado em vários problemas da literatura e seus resultados comparados à técnica míope de subconjuntos aninhados denominada RFE-SVM. Uma importante contribuição teórica do trabalho se refere ao desenvolvimento do conceito de margem projetada. A utilização deste valor, computado como a projeção da margem real de um espaço em cada subespaço de dimensão inferior, associado ao algoritmo IMA, permitiu maior eficiência e rapidez na solução dos problemas de classificação e, portanto, no processo de busca como um todo.

Palavras-chave – Seleção de Características, Classificador de Larga Margem, Busca Ordenada, Margem Projetada.

Abstract – Support Vector Machines have been widely used quite effectively on high dimensional problems. In this sense, it is important to develop new feature selection strategies that are associated with this type of classifier. In this paper, we propose a new method for feature selection based on an ordered search process, also known as best-first, to explore the space of possible candidates. The algorithm, called AOS, uses as a search measure the margin values calculated using a large margin classifier, called IMA. This highly efficient classifier allows great flexibility and speed in obtaining the margin values, enabling the solution of problems of reasonable size, with hundreds of features, and avoiding combinatorial explosion. The algorithm was tested on several problems from the literature and the results were compared to the RFE-SVM. An important theoretical contribution of the paper refers to the concept of the projected margin. This value, computed as the projection of the maximal margin vector on a lower dimensional subspace, is used as an upper bound to the actual maximal margin. This enables greater efficiency and speed in solving problems of classification and, therefore, in the search process as a whole.

Keywords – Feature Selection, Large Margin Classifier, Ordered Search, Projected Margin.

1. INTRODUÇÃO

O objetivo principal do processo de seleção de características é a eliminação de variáveis irrelevantes com o intuito de produzir subconjuntos de variáveis relevantes que sejam capazes de generalizarem melhor para um dado problema de classificação. Também, podem-se destacar como importantes questões relativas ao requerimento de tempo de computação, descoberta de variáveis que têm maior poder discriminante, como em análise de genes, bem como uma melhor visualização e interpretação dos resultados. Neste sentido, considera-se neste trabalho a investigação da eficiência da utilização das máquinas de vetores suportes associadas a um processo ordenado de seleção de candidatos na obtenção dos subconjuntos com maior poder de generalização. Adota-se, para tanto, uma estratégia de solução do tipo reversa na qual as variáveis com menor poder de discriminação são retiradas do problema. Ao contrário do algoritmo RFE, que retira uma variável por vez de forma irrevogável definindo uma sequência de subconjuntos aninhados, emprega-se um processo de busca ordenada que gera uma árvore de possibilidades e permite uma maior exploração da interdependência entre o conjunto de variáveis do problema. Como forma de evitar a explosão combinatória decorrente do número exponencial de possibilidades, utilizam-se de duas estratégias que permitem controlar o processo de busca. Primeiramente, a quantidade de variáveis é reduzida até um tamanho gerenciável com a utilização de uma técnica míope como

a eliminação de variáveis sugerida pelo algoritmo RFE ou por métodos de filtragem que se baseiam no estabelecimento de um ranking de variáveis segundo medidas obtidas por critérios estatísticos ou de informação. Neste trabalho, emprega-se com este objetivo a utilização de um classificador SVM que minimiza em seu treinamento a norma L_1 do vetor w no sentido de obter soluções mais esparsas. Em segundo lugar, limita-se o fator de ramificação com a retirada de duas, ou no máximo de três, possíveis variáveis a cada nível da árvore de busca. Exaustivos testes demonstraram que, em um processo de eliminação reverso, a adoção de um fator de ramificação maior não altera de forma significativa a escolha dos melhores subconjuntos de variáveis. O trabalho está organizado como se segue. Nos capítulos 2 e 3, descreve-se de forma sucinta alguns tópicos preliminares relacionados respectivamente ao problema de classificação e a base teórica do algoritmo IMA. Em seguida, no capítulo 4, descreve-se o problema de seleção de características e a base teórica dos algoritmos RFE e AOS. No capítulo 5, descreve-se o critério estabelecido para a retirada de variáveis, denominado critério de ramificação, e medidas de avaliação, como exemplo, as margens real e projetada. Finalmente, no capítulo 6, são apresentados os experimentos realizados e os resultados obtidos.

2. CLASSIFICAÇÃO

2.1. Problema de Classificação Binária

Seja um conjunto de dados Z de cardinalidade m, denominado conjunto de treinamento, composto de um conjunto de vetores x_k e de um conjunto de escalares y_k. Cada vetor, rotulado por um valor escalar, está inserido em um espaço de dimensão d, $x_k \in R^d$, chamado de espaço de entrada do problema, representando uma respectiva amostra ou exemplo. Considerando que o valor de cada escalar y_k representa a classe de cada vetor x_k, tem-se para problemas de classificação binária, $y_k \in \{-1, +1\}$ para $k = \{1, \cdots, m\}$. Para problemas linearmente separáveis, um classificador linear será representado no espaço de entrada por um hiperplano, dado pela seguinte equação:

$$f(x) = \langle w, x \rangle + b, \tag{1}$$

onde w representa o vetor normal ao hiperplano e b o valor do viés (*bias*).

Considerando a integração do viés da equação em uma componente adicional do vetor w, adicionando, também, uma componente +1 no vetor representativo de cada ponto, tem-se a representação do classificador, no espaço denominado $\Phi-$*space*, na forma:

$$f(x) = \langle w, \Phi(x) \rangle \tag{2}$$

A resposta do classificador poderá ser obtida através da aplicação de uma função sinal φ ao valor do discriminante relacionado à equação do hiperplano, ou seja:

$$\varphi(f(x)) = +1 \text{ se } f(x) \geq 0 \text{ ou } \varphi(f(x)) = -1 \text{ se } f(x) < 0 \tag{3}$$

2.2. Algoritmo Perceptron

O algoritmo desenvolvido por Rosenblatt [1] pode ser utilizado para a determinação do vetor w em um número limitado de iterações. A quantidade de iterações está relacionada à quantidade de atualizações do vetor de pesos e, consequentemente, à quantidade de erros cometidos pelo algoritmo.

Para uma determinada amostra do conjunto de treinamento, ocorrerá um erro ou uma classificação incorreta se:

$$y_k(\langle w, \Phi(x_k) \rangle) < 0 \tag{4}$$

Neste sentido, pode-se adotar como função de perda a quantidade de amostras classificadas incorretamente. Esta função, definida como a função de perda 0-1, é descrita como:

$$J(w) = \sum_k 1 | \{\varphi(f(x_k)) \neq y_k\} \tag{5}$$

ou

$$J(w) = \sum_k \text{Max}\{0, \varphi(-y_k(\langle w, \Phi(x_k) \rangle))\}, (x_k, y_k) \in Z \tag{6}$$

Entretanto, sendo esta função constante por partes e, portanto, não diferenciável, torna-se mais apropriado à utilização de uma nova função de perda, linear por partes, dada pela soma negativa de todos valores funcionais, também chamados de valores de margens, das amostras classificadas incorretamente. Ou seja:

$$J(w) = \sum_k \text{Max}\{0, -y_k(\langle w, \Phi(x_k) \rangle)\}, (x_k, y_k) \in Z, \tag{7}$$

tornando possível a utilização do método do gradiente.

Portanto, caso o problema seja linearmente separável no $\Phi-$*space*, para se determinar uma solução que minimize a função de perda em relação ao vetor w, é necessário avaliar o vetor gradiente considerando, somente, a ocorrência das amostras classificadas incorretamente. Este processo, aplicado individualmente a uma amostra, resulta na seguinte regra de correção:

$$w_{(t+1)} = w_{(t)} + \eta \cdot \Phi(x_k) \cdot y_k, (x_k, y_k) \in Z, \tag{8}$$

sendo η a taxa de aprendizado.

2.3. Representação Dual

Para obter-se os pontos suportes relacionados à descrição da equação do hiperplano separador é necessário resolver o problema de inequações relacionado ao algoritmo de treinamento do Perceptron na sua forma dual. Para tanto, é preciso representar o vetor w como uma combinação linear positiva dos pontos do conjunto de treinamento. Tal expansão define um conjunto de escalares positivos, chamados de multiplicadores ou variáveis duais, sendo representados pelo vetor α, $\alpha \in R^m$. Portanto:

$$w = \sum_k \alpha_k y_k \Phi(x_k), \alpha_k \geq 0 \tag{9}$$

Substituindo esta representação na equação original, tem-se uma nova forma da função:

$$f(x_i) = \sum_k \alpha_k y_k y_i \langle \Phi(x_k), \Phi(x_i) \rangle \tag{10}$$

Neste sentido, para trabalhar-se somente com o conjunto de variáveis duais α, é necessário reescrever a regra de correção do Perceptron na sua forma dual. Para a forma dual do algoritmo, pode-se fazer a mesma atualização, considerando a nova descrição do vetor w:

$$w = \sum_k \alpha_k y_k \Phi(x_k) + \eta \cdot y_i \Phi(x_i), \tag{11}$$

resultando, consequentemente, para uma amostra classificada incorretamente, na atualização do respectivo multiplicador com base na expressão:

$$\alpha_k = \alpha_k + \eta \cdot 1 \tag{12}$$

Para a atualização do viés, representado como uma componente adicional do vetor w, tem-se que considerar o fato de que cada vetor x_k possui um valor adicional $+1$ na sua última componente. Neste sentido, o valor do viés pode ser computado separadamente em um esquema do tipo *online* conforme o vetor de pesos.

Esta representação dual do modelo Perceptron é também chamada de representação dependente dos dados, podendo ser interpretada como um classificador *kernel*. A medida de similaridade entre os dados é computada pelo produto interno dos vetores do espaço de entrada ou dos vetores característicos se for considerada a existência do Φ-*space*.

2.4. Percetron de Margem Fixa

Leite e Fonseca [2] propõem uma nova formulação para o modelo perceptron, no sentido de garantir que o conjunto de exemplos guarde uma distância mínima em relação ao hiperplano separador sem limitar diretamente o valor da norma do vetor w. Para tanto, é considerada a restrição de que cada amostra deva possuir um valor de margem geométrica correspondente superior ou igual ao valor estabelecido como margem fixa, sendo o valor da margem geométrica definido como o valor da margem funcional da respectiva amostra dividido pelo valor da norma euclidiana do vetor w. Isto equivale à realização do produto interno do vetor $\Phi(x_k)$ pelo vetor unitário de direção w, representado por $w/\|w\|_2$. Assim, deve-se resolver o seguinte sistema de inequações não lineares para determinado valor de margem fixa representado pelo parâmetro γ_f:

$$y_k(\langle w, \Phi(x_k) \rangle)/\|w\|_2 \geq \gamma_f \tag{13}$$

Em função desta modificação, torna-se necessário reescrever a função de perda do modelo de forma a possibilitar a obtenção de uma nova regra de correção. A nova função será equivalente à soma dos valores das respectivas margens geométricas dos exemplos que erram, considerando o desconto do valor da margem fixa. Ou seja:

$$J(w) = \sum_k \text{Max}\{0, \gamma_f - y_k(\langle w, \Phi(x_k) \rangle)/\|w\|_2\}, (x_k, y_k) \in Z \tag{14}$$

Portanto, ao contrário do algoritmo básico do perceptron, considera-se também como erro aqueles exemplos que, embora classificados corretamente, não estejam a uma distância mínima, no sentido geométrico, do hiperplano separador. A solução do sistema de inequações pode ser considerada como aquela que minimiza a função de erro J. Neste sentido, tomando-se o gradiente da função em relação ao vetor w, tem-se a seguinte correção, caso ocorra um erro, $y_k(\langle w, \Phi(x_k) \rangle) < \gamma_f \cdot \|w\|_2$, aplicada a uma determinada amostra $(x_k, y_k) \in Z$:

$$w_{(t+1)} = w_{(t)} - \eta(\gamma_f \cdot w/\|w\|_2 - y_k \cdot \Phi(x_k)) \tag{15}$$

3. ALGORITMO DE MARGEM INCREMENTAL

3.1. Formulação de Máxima Margem

Uma nova formulação para o problema de maximização da margem foi desenvolvida [2] a partir de duas constatações importantes. Primeiramente, observando o fato de que na obtenção da máxima margem, os pontos ou vetores suportes de classes contrárias se encontram à uma mesma distância do hiperplano separador. Ou seja, considerando as margens das classes de rótulos positivo e negativo, tem-se $\gamma^+ = \gamma^-$, onde:

$$\begin{aligned} \gamma^+ &= \text{Min } y_k \cdot f(x_k), \text{ para todo } x_k \in X^+ \\ \gamma^- &= \text{Min } y_k \cdot f(x_k), \text{ para todo } x_k \in X^- \end{aligned} \tag{16}$$

Em segundo lugar, observando a possibilidade da obtenção de soluções de larga margem, em um número finito de correções, na solução do problema do perceptron de margem geométrica fixa, na forma:

$$y_k \cdot f(\Phi(x_k)) \geq \gamma_f \cdot \|w\|_2, \text{ para valores de } \gamma_f < \gamma^* \tag{17}$$

Neste sentido, propõe-se a solução aproximada do problema de máxima margem, considerando a maximização explícita e direta da margem geométrica. Neste contexto, deve-se resolver o seguinte problema de otimização:

$$\begin{aligned} &\text{Max}_w \gamma_g \\ &\text{Sujeito a} \\ y_k \cdot f(\Phi(x_k)) &\geq \gamma_g \cdot \|w\|_2, k = 1, \cdots, m \end{aligned} \tag{18}$$

Considerando o fato de que a regularização do vetor w está implícita na função de perda apresentada na seção 2.2, ocorre uma limitação no crescimento do valor da norma, impedindo que a mesma escape para valores muito altos. Assim, torna-se possível computar diretamente o valor da maior margem geométrica, a qual se aproxima suficientemente da margem ótima no sentido de garantir a construção de um classificador de larga margem.

3.2. Norma Arbitrária

Uma importante flexibilidade fornecida pelo Algoritmo de Margem Incremental, IMA$_p$ [3], está no fato de poder-se trabalhar livremente com qualquer norma diferenciável relacionada ao vetor w. Para tanto, para diferentes valores de margens fixas, considera-se a solução de sucessivos problemas na forma:

$$y_k \cdot f(\Phi(x_k)) \geq \gamma_f \cdot \|w\|_q, k = 1, \cdots, m, \tag{19}$$

no sentido de minimizar a norma L_q, dada por $\|w\|_q = (\sum_k |w_k|_q)^{1/q}$, do vetor w e de se estabelecer uma solução de margem L_p, baseando-se no fato de que as normas conjugadas p e q satisfazem a relação: $1/p + 1/q = 1$.

Para tanto, torna-se necessário definir a função de perda do modelo perceptron de margem fixa em sua forma geral, relacionada a existências de uma norma q conjugada, ou seja:

$$J(w) = \sum_k \text{Max}\{0, \gamma_f - y_k(\langle w, \Phi(x_k)\rangle)/\|w\|_q\}, (x_k, y_k) \in Z \tag{20}$$

Neste sentido, tomando-se a derivada da função em relação ao vetor w, tem-se a seguinte expressão que define o vetor gradiente local:

$$\nabla_w J(w) = ((\gamma_f \cdot \varphi(w_k) \cdot |w_k|^{q-1})/\|w\|_q^{q-1}) - y_k \cdot \Phi(x_i) \tag{21}$$

Caso ocorra um erro, $y_k(\langle w, \Phi(x_k)\rangle) < \gamma_f \cdot \|w\|_q$, a seguinte regra de correção será aplicada a uma determinada amostra $(x_k, y_k) \in M$:

$$w_{(t+1)} = w_{(t)} - \eta((\gamma_f \cdot \varphi(w_k) \cdot |w_k|^{q-1})/\|w\|_q^{q-1} - y_k \cdot \Phi(x_k)) \tag{22}$$

Esta nova solução verifica-se como oportuna na medida em que podem ser utilizados quaisquer valores de norma variando entre L_1 e L_∞ preservando-se a mesma estrutura do Algoritmo de Margem Incremental e modificando-se somente a equação de correção do vetor w.

A minimização da norma L_1 do vetor w define um hiperplano separador com margem L_∞. Ou seja, a distância computada dos pontos ao hiperplano separador é tal que maximiza o valor da maior componente do vetor normal. Esta variante é aconselhável se for necessário a obtenção de soluções mais esparsas em relação às componentes do vetor w como no processo de seleção de características. Neste caso, é possível constatar que a solução ou hiperplano proveniente da formulação L_∞, se posiciona sempre quase perpendicular ao eixo da maior componente, tornando-se dependente desta coordenada.

4. SELEÇÃO DE CARACTERÍSTICAS

O processo de seleção de características (*feature selection*), ou seleção de variãveis, se baseia na seleção, segundo algum critério, de um subconjunto do conjunto original de características do problema que produza os mesmos (ou quase mesmos) resultados. A técnica é aplicada nas mais diversas áreas como, por exemplo, reconhecimento de padrões, mineração de dados e aprendizado de máquina. Dentre algumas aplicações em problemas reais, pode-se destacar a categorização de textos, recuperação de imagens, detecção de intrusos e análise genômica.

4.1. Seleção por Filtro

Essa abordagem de seleção introduz um processo separado, o qual ocorre antes da aplicação do algoritmo de aprendizado. O modelo foi batizado de filtro [4] pelo fato de "filtrar" os atributos irrelevantes, segundo algum critério, antes que uma indução ocorra. Sendo assim, métodos de filtros são independentes do algoritmo de classificação que, simplesmente, receberá como entrada o conjunto de exemplos contendo apenas os atributos selecionados pelo filtro. A vantagem do modelo em filtro está no fato de que o mesmo não precisa ser reaplicado para cada execução do algoritmo de treinamento. Contudo, sabe-se, que na maioria das vezes, o subconjunto ótimo de características depende do algoritmo de treinamento a ele associado. Assim, um subconjunto de características selecionado usando um método em filtro pode resultar em uma alta precisão para determinado classificador e em baixa precisão em outros. O maior problema destes métodos é que cada coeficiente é computado utilizando somente informações do atributo relacionado, não levando em conta a existência de uma interdependência, ou mútua informação, entre as características. De fato, podem existir atributos complementares que individualmente não têm uma relevância, mas que combinados podem ter um papel importante no processo de discriminação. Dentre os métodos de filtro, pode-se citar o Gulob [5], método este, que se baseia na diferença da magnitude dos níveis de expressão de cada atributo. Um atributo pode ser dito diferencialmente expresso em duas classes distintas se a diferença da média ponderada dessas classes dividida pela soma do seu desvio padrão for elevada. Definindo μ_1 e μ_2 como as médias das duas classes, e σ_1 e σ_2 como seus desvios padrões, respectivamente, pode-se definir as significâncias (importâncias) Golub G da seguinte forma:

$$G = \frac{|\mu_1 - \mu_2|}{(\sigma_1 + \sigma_2)} \tag{23}$$

4.2. Seleção Embutida

A estratégia dos métodos embutidos, também chamada de seleção em cápsula, se baseia no fato de que alguns indutores são capazes de realizar sua própria seleção de atributos. Eles fazem uso do algoritmo de indução para estimar o valor do subconjunto de características selecionado durante a fase de treinamento. A idéia central destes métodos se baseia na otimização direta de uma função objetiva composta geralmente de dois termos. O primeiro, relacionado à uma medida do desempenho do classificador, que deve ser maximizada, e o segundo, uma medida de regularização relacionada à quantidade de variáveis, que deve ser minimizada. Um exemplo de método embutido é fazer a utilização do algoritmo IMA$_p$ na sua formulação L_∞, que minimiza a norma L_1 do vetor w. Nessa versão, ele seleciona as características pelo valor da maior componente do vetor, minimizando as componentes não importantes, sendo essas, assim, facilmente eliminadas.

4.3. Seleção *Wrapper*

Em contraste com filtros, a abordagem *wrapper* gera vários subconjuntos de atributos como candidatos, executa o indutor individualmente em cada subconjunto e usa a precisão do classificador para avaliar o subconjunto em questão. Este processo é repetido até que um critério de parada seja satisfeito. A idéia geral desta abordagem é que o algoritmo de seleção de características existe como um *wrapper* ao redor do indutor e é responsável por conduzir a busca por um bom subconjunto de atributos.

Em geral, a busca é conduzida no espaço do subconjunto de atributos, ou de candidatos. Como estratégia pode-se empregar algoritmos míopes (*greedy*) ou buscas direcionadas (*best-first*) e com direções *forward* ou *backward*. A precisão dos subconjuntos pode ser estimada por uma validação cruzada (*cross-validation*). A principal vantagem deste modelo é a dependência entre os algoritmos de seleção e de aprendizado. Por outro lado, essa abordagem pode ser computacionalmente dispendiosa, uma vez que o indutor deve ser executado para cada subconjunto de atributos considerado.

4.3.1. *Recursive Feature Elimination*

A idéia básica do método da eliminação recursiva de características, denominado de *Recursive Feature Elimination* (RFE) [5], é a eliminação recursiva da menor componente do vetor w, uma vez que ela não tem muita influência sobre a posição do hiperplano. A cada passo do processo, um número fixo de componentes é eliminado e o classificador é retreinado. A eliminação recursiva de uma característica por vez gera um classificador com menos erros esperados, quando comparada à remoção de mais de uma característica ao mesmo tempo. Uma consideração a ser feita em relação ao método RFE é que, se a margem geométrica do vetor w for definida por:

$$\gamma_g = \frac{\gamma_g}{\|w\|} w, \tag{24}$$

pode-se observar que a magnitude das componentes $|w_j|$ do vetor tem uma relação direta com a magnitude da margem projetada do vetor em um subespaço aonde a j-ésima característica é excluída, como é dada pela equação:

$$\gamma_{pj}^{d-1} = \frac{\gamma_g^d}{\|w\|}(\sum_{k \neq j} w_k^2)^{\frac{1}{2}}, \tag{25}$$

onde γ_{pj}^{d-1} indica a magnitude da margem geométrica quando removida a característica j em um espaço de dimensão d. Por razões de simplicidade, chama-se este valor de margem projetada.

É de fácil constatação que a remoção da menor componente relacionada com a menor margem projetada não resultará sempre na melhor margem geométrica para um problema de dimensão $d-1$. Nesse sentido, se estabelece uma relação direta, usando a Minimização do Risco Estrutural, entre o valor da margem e a capacidade de generalização do classificador, e percebe-se que o RFE nem sempre escolhe a característica adequada para remover. Assim, o uso de uma estratégia míope de exploração dos espaços possíveis não garante sempre um classificador ótimo associado ao processo de seleção de características.

4.3.2. *Admissible Ordered Search*

Técnicas de seleção de características comumente utilizadas com base em análises de variância dos dados, bem como métodos baseados na eliminação recursiva de características, nem sempre encontram classificadores com uma menor quantidade de atributos ou um melhor poder de generalização. Neste sentido, foi introduzido um novo algoritmo de seleção de características, denominado *Admissible Ordered Search* (AOS), o qual se baseia na realização de uma busca ordenada admissível. Este algoritmo tem a capacidade de encontrar, em cada dimensão do problema, o classificador de maior margem.

Em um processo de busca ordenada, garante-se a admissibilidade do algoritmo se a função de avaliação é monótona. Para problemas de minimização, esta função precisa ser monótona crescente, e, para problemas de maximização, esta função precisa ser monótona decrescente. Neste sentido, uma vez que se está à procura da maximização da margem, usa-se como função de avaliação o valor da margem obtido a partir de cada hipótese após a solução do problema de classificação no conjunto de características selecionado por essa hipótese. A admissibilidade do processo é garantida, uma vez que os valores da margem são sempre decrescentes quando a dimensão diminui. Ou seja:

$$\gamma_{gj}^{d-1} \leq \gamma_g^d, \forall j \tag{26}$$

A estratégia de controle dessa busca ordenada é implementada com a inserção das hipóteses candidatas em uma fila ordenada pelos valores das margens. Uma vez que a ordem das características selecionadas não importa, poderá haver alguma redundância. A fim de evitar este problema, cria-se uma tabela hash, e, para cada nova hipótese, verifica-se a sua unicidade nessa tabela antes de inseri-la na fila. Usando a margem geométrica real exigiria a solução de um problema para a maximização da margem para cada hipótese gerada. Em vez disso, utiliza-se uma estimativa otimista desta margem, que é a sua margem projetada. Este valor será um limite máximo para a margem geométrica real da mesma hipótese no espaço associado. Com isso, mantém-se a admissibilidade do processo, uma vez que:

$$\gamma_{pj}^{d-1} \geq \gamma_{gj}^{d-1}, \forall j \tag{27}$$

$$\gamma_{pj}^{d-1} \leq \gamma_g^d, \forall j \tag{28}$$

Para cada iteração do algoritmo, a hipótese relacionada com a maior margem é escolhida, independentemente da sua dimensão, para ser expandida e gerar novas hipóteses num espaço de dimensão menor. Desta forma, pode-se verificar duas situações possíveis:

- Primeiro: para os casos em que o valor da margem para a hipótese escolhida é o valor projetado, pode-se calcular o seu valor real através da solução de um problema de maximização da margem e compará-lo com o maior valor da fila de prioridade. Se ainda é a melhor opção, fecha-se este estado e gera-se as suas hipóteses. Caso contrário, substitui-se o valor da margem projetada dessa hipótese pelo valor real e reinsere o mesmo na fila;
- Segundo: para os casos em que o valor da margem da hipótese escolhida já é o valor real, fecha-se o estado e gera-se suas hipóteses.

Para amenizar a explosão combinatória, o algoritmo desenvolve um esquema de poda adaptativa, baseado na atualização constante de um limite de margem inferior. Este limite é computado toda vez que uma hipótese escolhida for a primeira a chegar em uma dada dimensão. Para tanto, é utilizada uma estratégia de seleção míope que avalia os valores de margem até uma dimensão inferior escolhida, eliminando, assim, estados de maior dimensão que têm valores inferiores a este limite. Também, neste momento, uma estimativa do erro esperado pode ser calculada e utilizada como critério de parada do algoritmo, uma vez que representa o desempenho de generalização do classificador.

5. CRITÉRIOS DE SELEÇÃO ASSOCIADOS AO CLASSIFICADOR DE LARGA MARGEM

5.1. Ramificação

Como parâmetro de ramificação, adota-se a eliminação das componentes relacionados aos menores valores de margens projetadas, ou as menores componentes do vetor w, tal como utilizado no RFE. Este critério está relacionado à escolha da menor variação no valor da função objetiva do problema SVM de minimização em sua formulação dual:

$$J = 1/2 \cdot \alpha^T \cdot H \cdot \alpha - \alpha^T \cdot 1, \tag{29}$$

sendo α o vetor de multiplicadores e H a matriz definida a partir da matriz *kernel* com componentes na forma $y_i \cdot y_j \cdot K_{i,j}$. Assim, sustentando o mesmo vetor de multiplicadores, se for retirada a i-ésima variável, tem-se a variação:

$$\Delta J(i) = J - J(i) = 1/2 \cdot \alpha^T \cdot H \cdot \alpha - 1/2 \cdot \alpha^T \cdot H(i)\alpha = 1/2 \cdot \alpha^T \cdot (H - H(i)) \cdot \alpha \tag{30}$$

Desta forma escolhe-se como variável a ser retirada aquela que produzir a menor variação no valor da função, ou seja:

$$k = \text{Arg Min}_i \Delta J(i) \tag{31}$$

Em sua formulação primal, pode-se associar a variação do valor da função à variação da norma do vetor w, ou seja:

$$\Delta J(i) = J - J(i) = 1/2 \cdot \alpha^T \cdot H \cdot \alpha - 1/2 \cdot \alpha^T \cdot H(i) \cdot \alpha = \|w\|_2 - \|w_i\|_2 \tag{32}$$

O que se torna equivalente, em termos de critério, à escolha da variável associada a componente de menor magnitude do vetor. Ou seja:

$$k = \text{Arg Min}_i (w_i)^2 \tag{33}$$

5.2. Solução Inicial

Como foi visto, a escolha da variável associada à componente de menor magnitude do vetor w está relacionada à escolha da margem projetada de maior valor. De fato, se for escolhida uma variável associada a uma componente de valor zero, tem-se o valor da margem projetada igual ao valor da margem máxima real no subespaço associado. Isto facilita bastante o processo de solução dos problemas de classificação com SVM, pois, neste caso, não há necessidade de se resolver o novo problema de classificação, uma vez que as soluções são as mesmas.

Quando o valor da componente de menor magnitude for muito baixo, tem-se a solução dos dois problemas muito próximas. Neste caso, utiliza-se a última solução obtida no espaço em questão como solução inicial do problema de classificação do subespaço inferior associado. Se a formulação for primal elimina-se a componente associada do vetor w. Entretanto, caso a formulação seja dual utiliza-se o mesmo vetor de multiplicadores α.

Para o cálculo da margem projetada na formulação dual utiliza-se a mesma relação de normas do vetor w computados a partir de sua expansão em função do vetor de multiplicadores α. Ou seja:

$$\gamma_{pj}^{d-1} = \left(\frac{1}{\|w\|_2}\right) \cdot \gamma_g^d \cdot \|w_j\|_2 = \left(\frac{1}{\alpha^T H \alpha}\right) \cdot \gamma_g^d \cdot \alpha^T H(j) \alpha \tag{34}$$

6. EXPERIMENTOS E RESULTADOS

6.1. Conjuntos de Dados

Foram utilizadas seis bases de dados para análise dos resultados. Quatro bases linearmente separáveis, sendo duas sintéticas e duas de *microarrays*, e duas não linearmente separáveis. Todas as bases usadas neste trabalho estão contidas no repositório de aprendizado de máquinas da UCI [6].

A base Synthetic possui 600 exemplos de 6 tipos de gráficos de controle gerados sinteticamente. Ela possui 60 componentes e foi adaptada para um problema de classificação binária e linearmente separável. A base Robot LP4 possui 117 instâncias e 90 atributos, com as classes divididas entre normal versus colisão e obstrução.

As bases de *microarrays* são as bases Leukemia e Breast. A base de dados Leukemia visa classificar pacientes com leucemia por meio de sua expressão genética. Os dados são uma combinação de amostras de treinamento e validação de 47 pacientes com leucemia linfóide aguda e 25 pacientes com leucemia mielóide aguda. As amostras apresentam informações referentes a 7129 genes. A base Breast contém biópsias de 24 pacientes de câncer de mama antes dos 4 ciclos de tratamento taxotere (docetaxel). Ela conta com 12625 atributos (genes).

As bases não linearmente separáveis são a Sonar e a Ionosphere. Na realidade, a base Sonar é linearmente separável, porém a retirada de qualquer característica a deixa não linear. Por isso, ela foi tratada como não linear. Ela é constituída por classificação de sinais sonares. Ela possui 60 componentes e 208 amostras. Já a base Ionosphere descreve dados sobre radares. Bons resultados são considerados se mostram evidência de algum tipo de estrutura na ionosfera, caso contrário são considerados ruins. Esse conjunto de dados é composto por 34 atributos e 351 exemplos.

6.2. Resultados

Os resultados são mostrados na tabela 1. A coluna SVM corresponde à base completa, com o conjunto total de atributos. As comparações foram feitas entre os algoritmos Golub, RFE e AOS. Para o AOS existem 2 colunas, uma referente à comparação com a mesma quantidade de atributos que o RFE atingiu e outra para a quantidade final que o AOS conseguiu atingir. Para as bases não linearmente separáveis foi utilizado um *kernel* gaussiano com o parâmetro igual a 1. Como medidas de qualidade, foram apresentados a margem obtida para cada solução, além do erro estimado utilizando uma validação cruzada com um 3-fold.

Tabela 1: Resultados das bases

—	SMV	Golub-SVM	RFE-SVM		AOS-SVM		AOS-SVM (Final)	
			L_1	L_2	L_1	L_2	L_1	L_2
Synthetic								
Atributos	60	38	7	8	7	8	6	6
Margem	6,497	0,415	0,523	0,482	0,574	1,189	0,132	0,132
Erro 3-fold	0,17%	4,33%	0,83%	1,00%	0,50%	0,17%	1,17%	1,17%
LP4 Robot								
Atributos	90	72	7	11	7	11	5	8
Margem	6,067	5,363	0,922	1,631	0,922	1,631	0,473	0,279
Erro 3-fold	11,97%	11,97%	2,56%	5,13%	2,56%	5,13%	2,56%	5,98%
Leukemia								
Atributos	7129	8	5	5	5	5	3	3
Margem	13444,160	425,495	2257,015	2257,015	2257,015	2257,015	354,366	354,366
Erro 3-fold	5,56%	9,72%	2,78%	2,78%	2,78%	2,78%	1,39%	1,39%
Breast								
Atributos	12625	3	4	5	4	5	2	2
Margem	4830,087	15,842	401,127	621,436	401,127	668,616	258,508	258,508
Erro 3-fold	20,83%	20,83%	4,17%	12,50%	4,17%	4,17%	4,17%	4,17%
Sonar								
Atributos	60	25	7		7		6	
Margem	0,077	0,024	0,010		0,013		0,007	
Erro 3-fold	15,90%	25,01%	15,74%		14,90%		20,18%	
Ionosphere								
Atributos	34	10	7		7		5	
Margem	0,077	0,031	0,031		0,044		0,011	
Erro 3-fold	7,12%	10,26%	7,69%		6,84%		8,26%	

7. CONCLUSÃO

Neste trabalho, introduziu-se um novo algoritmo para a seleção de características, denominado AOS, que utiliza critérios e medidas de qualidade provenientes de um classificador de larga margem e explora com eficiência o espaço de possibilidades. Como pode ser observado, este algoritmo apresentou resultados bastante satisfatórios. Ele foi superior ao RFE em todos os experimentos. Na solução de mesma dimensão, sempre obteve margens iguais ou superiores e erros de generalização iguais ou inferiores. Já na solução final, conseguiu produzir sempre classificadores com um menor número de atributos e com um bom poder de generalização.

REFERÊNCIAS

[1] F. Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain". *Psychological Review*, vol. 65, pp. 386–408, 1958.

[2] S. C. Leite and R. F. Neto. "Incremental Margin Algorithm". *Neurocomputing*, vol. 71, pp. 1550–1560, 2007.

[3] S. M. Villela, R. F. Neto, S. C. Leite and A. E. Xavier. "Classificador de Máxima Margem com Norma Arbitrária: Formulação, Algoritmo e Resultados". In *IX CBRN - Congresso Brasileiro de Redes Neurais / Inteligência Computacional*, Ouro Preto, MG, 2009.

[4] G. H. John, R. Kohavi and K. Pfleger. "Irrelevant Features and the Subset Selection Problem". In *Proceedings of the Eleventh International Conference on Machine learning*, pp. 121–129, New Brunswick, NJ, 1994.

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

[6] A. Asuncion and D. J. Newman. *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science, 2007.