Algoritmo de margem incremental com norma p para classificadores de larga margem

Saulo Moraes Villela, Saul de Castro Leite, Raul Fonseca Neto
Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora
36036-330 Juiz de Fora, MG, Brasil
Email: {saulo.moraes,saul.leite,raulfonseca.neto}@ufjf.edu.br

***Resumo*—Este trabalho tem como objetivo apresentar uma nova formulação para o problema de obtenção de um classificador de larga margem com norma arbitrária. O trabalho apresenta, também, um novo algoritmo baseado na solução de sucessivos sistemas de inequações com o uso de uma versão modificada do algoritmo de treinamento do perceptron. A sua grande vantagem consiste na sua flexibilidade, que permite a utilização de diferentes valores de norma, ao contrário de algumas formulações especiais existentes que se restringem à utilização das normas L_1 e L_∞. O algoritmo foi chamado de IMA$_p$ pois trata-se de um caso especial do algoritmo IMA desenvolvido para a obtenção de classificadores de larga margem com o uso da norma euclidiana.**

***Index Terms*—Classificadores de larga margem, Norma p, Perceptron**

I. Introdução

Um dos principais problemas relacionados à teoria do aprendizado de máquinas consiste no aprendizado de uma função, ou hipótese, capaz de discriminar um conjunto de pontos, pertencentes a duas classes distintas, em um espaço real R^d, de dimensão d. Uma questão importante está associada à capacidade de generalização desse classificador. Neste sentido, tornou-se primordial a descoberta dos classificadores lineares de larga margem. O projeto de tais classificadores baseia-se no princípio da minimização do risco estrutural, proposto por Vapnik [1]. Em sua formulação original, esse problema foi desenvolvido no contexto da norma euclidiana L_2. Assim, torna-se interessante o desenvolvimento de novas formulações flexíveis para a solução desse tipo de problema que explorem valores diferentes de norma, bem como novos algoritmos de treinamento.

II. Problema de classificação linear com norma p

Seja $Z = \{z_i = (x_i, y_i) : i \in \{1, \cdots, m\}\}$ 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 linear 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:

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

Esse hiperplano pode não existir. Quando isso acontece, o conjunto Z é denominado não linearmente separável. Neste artigo assume-se que Z é linearmente separável, salvo indicação em contrário.

É dito que Z aceita uma margem $\gamma \geq 0$ quando existe um hiperplano $\mathcal{H} := \{x \in \mathbb{R}^d : w \cdot x + b = 0\}$ 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 [2]:

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

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

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

III. Perceptron de Margem Geométrica Fixa

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

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

A formulação desse problema, para a norma euclidiana, foi proposta em [3], sendo de modo similar ao Perceptron de Rosenblat [4], e o algoritmo resultante foi denominado de Perceptron de Margem Geométrica Fixa (*Geometric Fixed Margin Perceptron* – GFMP).

Para uma norma p, a função de erro, ou perda, $J : \mathbb{R}^{d+1} \to \mathbb{R}$ é definida como:

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

onde $\mathcal{M}$ é o subconjunto de Z que viola (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), $t \in \{1, 2, \dots\}$. Se o par z_i escolhido for um erro, ou seja, se $y_i (w^t \cdot x_i + b^t) < \gamma_f ||w^t||_q$, 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, $j \in \{1, \dots, d\}$, e à b quando $1 < q < \infty$, tem-se:

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

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

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

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

Para se ter uma regra de atualização para os casos onde $p = 1$ ou $p = \infty$, pega-se o limite $q \uparrow \infty$ e $q \downarrow 1$, respectivamente, na expressão da $\partial J / \partial w_j$. Para $p = \infty$, a análise é trivial, seguindo a regra derivada.

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

Para $p = 1$, reescreve-se $\partial J / \partial w_j$ como:

$$\frac{\partial J(w,b)}{\partial w_j} = \frac{|w_j|^{q-1}}{\sum_{i=1}^{d} |w_i|^q} \gamma_f ||w||_q \text{sinal}(w_j) - y_i x_{ij},$$

e considera-se o limite:

$$L := \lim_{q \to \infty} \frac{|w_j|^{q-1}}{\sum_{i=1}^{d} |w_i|^q}.$$

Nota-se que:

$$\frac{|w_j|^{q-1}}{d \max_i |w_i|^q} \leq \frac{|w_j|^{q-1}}{\sum_{i=1}^{d} |w_i|^q} \leq \frac{|w_j|^{q-1}}{\max_i |w_i|^q}.$$

Assim, para $|w_j| < \max_i |w_i|$, tem-se $L = 0$.

Agora, para tratar o caso onde $|w_j| = \max_i |w_i|$, seja $M := \max_i |w_i|$ e o conjunto de índices $N := \{k \in \{1, \dots, d\} : |w_k| = M\}$ e seja n a cardinalidade de N. Então, tem-se:

$$\begin{aligned}L &= \lim_{q \to \infty} \frac{M^{q-1}}{nM^q + \sum_{i \notin N} w_i|^q} \times \frac{\frac{1}{(d-n)M^q}}{\frac{1}{(d-n)M^q}} \\ &= \lim_{q \to \infty} \frac{1/(d-n)M^{-1}}{n/(d-n) + \frac{\sum_{i \notin N} |w_i|^q}{(d-n)M^q}} = \frac{1}{nM} = \frac{1}{n||w||_\infty},\end{aligned}$$

e, com isso, tem-se a regra de atualização:

$$w_j^{t+1} = \begin{cases} w_j^t - \eta \left(\gamma_f \text{sinal}(w_j^t)/n^t + y_i x_{ij} \right) & \text{se } |w_j| = ||w||_\infty \\ w_j^t - \eta \left(y_i x_{ij} \right) & \text{se } |w_j| < ||w||_\infty \end{cases}$$

onde n^t é a cardinalidade do conjunto $N^t := \{k \in \{1, \dots, d\} : |w_k^t| = ||w^t||_\infty\}$.

O algoritmo 1 descreve o Perceptron de Margem Geométrica Fixa com norma p.

Algoritmo 1: Perceptron de Margem Geométrica Fixa

Entrada: conjunto $Z = \{(x_i, y_i)\}, i \in \{1, \dots, m\}$;
margem geométrica fixa γ_f;
limite superior no número de iterações max;
Saída: vetor de pesos w e bias b;

início
  inicializar $\left(w^0, b^0\right)$;
  $j \leftarrow 0$;
  $t \leftarrow 0$;
  $stop \leftarrow$ falso;
  enquanto $j \leq max$ **e** $\neg stop$ **faça**
    $erro \leftarrow$ falso;
    para i **de** 1 **até** m **faça**
      se $y_i (w^t \cdot x_i + b^t) < \gamma_f ||w^t||_q$ **então**
        $w^{t+1} \leftarrow w^t - \eta \left(\gamma_f ||w^t||_q^{1-q} |w^t|^{q-1} \text{sinal}(w^t) + y_i x_i \right)$;
        $b^{t+1} \leftarrow b^t + \eta y_i$;
        $t \leftarrow t + 1$;
        $erro \leftarrow$ verdadeiro;
      fim se
    fim para
    se $\neg erro$ **então**
      $stop \leftarrow$ verdadeiro;
    fim se
    $j \leftarrow j + 1$;
  fim enquanto
fim

IV. Algoritmo de Margem Incremental

Uma nova formulação para o problema de maximização da margem foi proposta por Leite e Fonseca Neto [3] 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:

$$\begin{aligned}\gamma^+ &= \text{Min } y_i (w \cdot x_i + b), \text{ para todo } x_i \in Z^+ \\ \gamma^- &= \text{Min } y_i (w \cdot x_i + b), \text{ para todo } x_i \in Z^-\end{aligned}$$

Considerando a possibilidade da obtenção de soluções de margem fixa, em um número finito de correções, na solução do problema do GFMP:

$$y_i (w \cdot x_i + b) \geq \gamma_f \cdot ||w||_q, \text{ para valores de } \gamma_f < \gamma^*,$$

onde γ^* é a máxima margem, os autores propõem a formulação e solução aproximada do problema de máxima margem, considerando a maximização explícita e direta da margem geométrica. Adaptando-se para uma norma p, deve-se resolver o seguinte problema de otimização:

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

A técnica de solução desenvolvida consiste em uma estratégia de aprendizado incremental, através da qual são obtidas sucessivas soluções do problema do GFMP, 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, sucessivamente, o problema de inequações não lineares:

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

sendo cada solução equivalente à solução do problema do Perceptron de Margem Geométrica Fixa.

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:

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

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

Pode-se observar que, nesse caso, haverá sempre a garantia de solução do novo problema, já que a nova margem fixa é sempre inferior à margem ótima, ou seja, $\gamma_f^{t+1} = \left(\gamma^+ + \gamma^-\right)/2 < \gamma^*$. Tal condição deriva do fato de que se as margens negativa e positiva são desiguais, então a margem total não é máxima, implicando $\gamma^+ + \gamma^- < 2 \cdot \gamma^*$. Também se tem a garantia de que a nova margem fixa obtida é superior à margem fixa anterior, ou seja: $\gamma_f^{t+1} > \gamma_f^t$. Tal condição deriva-se da constatação das seguintes relações de exclusividade:

$$\gamma^+ > \gamma_f^t \text{ e } \gamma^- \geq \gamma_f^t \text{ ou } \gamma^+ \geq \gamma_f^t \text{ e } \gamma^- > \gamma_f^t,$$

garantindo um incremento no valor da nova margem fixa.

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:

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

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, a partir do qual, caso não haja uma nova solução do problema GFMP, adota-se como margem obtida o valor anterior da margem fixa, relacionado à última solução.

Uma forma adequada de escolha para essa constante é a de um valor proporcional à aproximação de margem desejada, isto é, se existe um desejo de uma aproximação α da margem ótima (ou seja, a margem final obtida pelo algoritmo é igual ou superior a $(1-\alpha)\gamma^*, \alpha \in (0,1)$), o valor de Δ deve ser escolhido como o valor de α. A prova disto é que, considerando um $t \geq 1$, tem-se $\gamma_f^t \in (0, \gamma^*)$ e $\gamma_f^{t+1} = (1+\alpha)\gamma_f^t$. Consequentemente, $\gamma_f^t \geq \gamma^*/(1+\alpha) > (1-\alpha)\gamma^*$. Sendo assim, adotando-se o último valor viável de margem γ^w, assegura-se que a última margem de parada é superior à aproximação desejada, ou seja, $\gamma^w \geq \gamma_f^t \geq (1-\alpha)\gamma^*$.

Adotando-se como solução inicial, ou ponto de partida, a solução do problema GFMP anterior, pode-se reduzir de forma gradativa a diferença existente entre o valor da margem final e as sucessivas margens fixas, tornando mais rápida a obtenção, a cada problema, do novo hiperplano separador.

O algoritmo de aprendizado (margem) incremental com norma p (*Incremental p-Margin Algorithm* – IMA$_p$) possui um *loop* principal relacionado às correções do valor da margem fixa e um procedimento aninhado relativo ao algoritmo de treinamento do GFMP. O Algoritmo 2 apresenta a descrição do mesmo considerando a chamada, a cada iteração, do algoritmo de treinamento do GFMP no sentido de fornecer uma solução viável até a obtenção do hiperplano separador final.

Algoritmo 2: Algoritmo de Margem Incremental

Entrada: conjunto $Z = \{(x_i, y_i)\}, i \in \{1, \ldots, m\}$;
limite superior no número de iterações max;
Saída: vetor de pesos w e bias b;
margem de parada γ^w;

início
  $\gamma_f \leftarrow 0$;
  repita
    $(w, b) \leftarrow$ GFMP(Z, γ_f, max);
    $\gamma^+ \leftarrow \text{Min}_{i^+} \left(w \cdot x_i + b \right)$;
    $\gamma^- \leftarrow \text{Min}_{i^-} \left(w \cdot x_i + b \right)$;
    $\gamma^w \leftarrow \text{Min} \left(\gamma^+, \gamma^- \right)$;
    se $\gamma^+ \neq \gamma^-$ **então**
      $\gamma_f \leftarrow \left(\gamma^+ + \gamma^- \right)/2$;
    senão
      $\gamma_f \leftarrow \left(1 + \Delta \right) \gamma_f$;
    fim se
  até *que a convergência do GFMP em max iterações não seja atingida*;
fim

Gentile [5] propôs um algoritmo on-line de larga margem, chamado ALMA$_p$. Internamente, o ALMA$_p$ utiliza uma generalização do perceptron denominada de algoritmo de norma p. O algoritmo funciona para o valores de norma $p \geq 2$. A regra de atualização é equivalente à do perceptron e seguida de um passo de normalização que controla o crescimento da norma. Essa normalização é necessária uma

vez que o algoritmo utiliza a margem funcional Γ_g ao invés da margem geométrica γ_g. Uma diferença importante entre os dois algoritmos é que o IMA$_p$ faz a atualização em um processo em *batch* após cada solução do GFMP, ao contrário do ALMA$_p$ que atualiza o valor da margem em um processo on-line após cada correção. Ao contrário do IMA$_p$, o ALMA$_p$ começa impondo um valor de margem alto e a cada iteração reduz o valor até atingir uma solução viável. Uma vantagem significativa da estratégia em *batch* do IMA$_p$ de gradualmente aumentar o valor da margem é que sempre se tem uma solução viável durante todo o processo.

V. Formulações Especiais

A. Formulação L_∞

Seja a minimização da norma L_1 do vetor w, definindo 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. Nesse 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 dessa coordenada. Tal afirmativa [6] pode ser evidenciada pelo fato de que a projeção L_∞, considerando somente a maior coordenada, pode desprezar as coordenadas com valor nulo sem afetar o valor da distância projetada.

Ainda segundo Rosset et. al. [6], é possível estabelecer o relacionamento entre os valores de margens de duas normas distintas pelas respectivas definições, ou seja, para o caso de escolha de L_1 e L_2:

$$\frac{y_i\left(w\cdot x_i+b\right)}{||w||_1}=\frac{y_i\left(w\cdot x_i+b\right)}{||w||_2}\cdot\frac{||w||_2}{||w||_1}$$

Pode-se, assim, observar que a margem L_∞, relacionada à margem conjugada, tende a ser maior quando a relação entre as normas for maior que a unidade, o que ocorre, geralmente, quando o vetor w for mais esparso, o que justifica o seu emprego no problema de seleção de características. Para a definição de um hiperplano separador com margem L_∞, pode-se optar pela solução do seguinte problema de otimização:

$$\text{Max}_w\text{Min}_{x_i\in X}\left\{y_i\left(w\cdot x_i+b\right)/||w||_1\right\}$$

Todavia, considerando o estabelecimento de uma margem funcional de valor unitário, ou seja:

$$\text{Min}_{x_i\in X}\left\{y_i\left(w\cdot x_i+b\right)\right\}=1,$$

obtém-se o seguinte problema equivalente:

$$\begin{gathered}\text{Min }||w||_1\\ \text{Sujeito a}\\ y_i\left(w\cdot x_i+b\right)\geq 1\end{gathered}$$

Substituindo o valor da norma L_1 do vetor w, tem-se:

$$\begin{gathered}\text{Min }\sum_j|w_j|\\ \text{Sujeito a}\\ y_i\left(w\cdot x_i+b\right)\geq 1\end{gathered}$$

Observando o fato de que $w=w^+-w^-$ implica em $|w_j|=w_j^++w_j^-$, Kecman e Hadzic [7] propõem a solução desse problema na forma:

$$\begin{gathered}\text{Min }\sum_j\left(w_j^++w_j^-\right)\\ \text{Sujeito a}\\ y_i\left\langle w^+-w^-,\Phi\left(x_i\right)\right\rangle\geq 1, w_j^+\geq 0 \text{ e } w_j^-\geq 0\end{gathered}$$

B. Formulação L_1

De outra forma, para a minimização da norma L_∞ do vetor w, definindo um hiperplano separador com margem L_1, ou seja, a distância dos pontos ao hiperplano separador é tal que maximiza o somatório em módulo das componentes do vetor normal, Pedroso e Murata [8] propõem uma formulação que culmina na solução de um único problema de Programação Linear. Considerando o fato de que: $||w||_\infty=\text{Max}_i|w_i|$, o problema de maximização da margem, segundo [8], pode ser reformulado como:

$$\text{Max}_w\text{Min}_{x_i\in X}\left\{y_i\left(w\cdot x_i+b\right)/\text{Max}_i|w_i|\right\}$$

Limitando o valor da margem funcional, de modo que $\text{Min}_{x_i\in X}\left\{y_i\cdot\left(w\cdot x_i+b\right)\right\}=1$, o problema de otimização se torna equivalente a:

$$\begin{gathered}\text{Max}_w\left\{1/\text{Max}_i|w_i|\right\}\\ \text{Sujeito a}\\ y_i\left(w\cdot x_i+b\right)\geq 1\end{gathered}$$

ou

$$\begin{gathered}\text{Min }z\\ \text{Sujeito a}\\ y_i\left(w\cdot x_i+b\right)\geq 1, z\geq+w_i \text{ e } z\geq-w_i\end{gathered}$$

Mesmo com a possibilidade de se resolver um único problema de Programação Linear, percebe-se que a condução do processo de otimização no espaço primal para as formulações L_1 e L_∞, torna inviável a solução de problemas de altíssima dimensão, já que a matriz de coeficientes possuirá posto relacionado à quantidade destes parâmetros. Nesse sentido, verifica-se como oportuna a utilização da formulação proposta na solução do problema de maximização da margem para quaisquer valores de norma.

VI. Experimentos e Resultados

A. Conjuntos de Dados

Para a análise dos resultados, foram utilizadas cinco bases linearmente separáveis, sendo três de *microarrays*. As bases utilizadas nesse artigo estão contidas no repositório de aprendizado de máquinas da UCI [9]. A base de dados Iris contém 3 séries de 50 instâncias das plantas Iris-Setosa, Iris-Versicolour e Iris-Virginica. Cada registro é composto por 4 atributos (larguras e comprimentos da pétala e da sépala). A combinação da classe Setosa contra as outras duas deixa a base linearmente separável. A base Mushroom contém atributos de diversas variedades de cogumelos e a tarefa do classificador é determinar quais são comestíveis. Ela conta com 98 componentes e possui 5644 amostras. A base Leukemia visa classificar pacientes com leucemia por meio de sua expressão genética. Os dados são

uma combinação de amostras de 47 pacientes com leucemia linfóide aguda e 25 pacientes com leucemia mielóide aguda e apresentam informações de 7.129 genes. A base Prostate contém expressões gênicas de câncer de próstata, com 12.600 genes, derivados de amostras de 52 pacientes com tumor e 50 com tecidos normais. A base Breast contém biópsias de 24 pacientes de câncer de mama antes dos 4 ciclos de tratamento taxotere. Ela conta com 12.625 genes.

B. Resultados

Foram comparados os valores de margem do IMA_p com as margens máximas encontradas pelas soluções de Programação Linear (PL), quando $p = 1$ ou $p = \infty$, e pelo SVM^{light} [10], para a norma euclidiana ($p = 2$). Os resultados são mostrados nas tabelas I, II e III. Os resultados do IMA_p estão divididos em alguns estágios, que são soluções do GFMP. Os números de iterações e atualizações estão acumulados para as chamadas ao GFMP. É possível observar que o algoritmo obtém soluções viáveis em um número pequeno de iterações e se aproxima das soluções exatas encontras pela PL e pelo SVM^{light}.

O algoritmo foi comparado, também, com o algoritmo ALMA_p. Para o ALMA_p, forma utilizados os parâmetros $B = 1/\alpha$ e $C = \sqrt{2}$, como sugerido por Gentile [5]. Para esses testes, as bases de dados foram normalizadas, sob uma norma p. Isso devido ao fato do ALMA_p só funcionar em bases normalizadas. As tabelas IV, V e VI mostram as comparações entre as execuções dos algoritmos ALMA_p e IMA_p para $p = 2$, $p = 6$ e $p = 10$, assim como feito em [5].

Tabela I
COMPARAÇÃO ENTRE O IMA_∞ E A SOLUÇÃO DE PL

Base	Algoritmo	GFMP	Iterações	Atualizações	Margem
Iris	IMA_∞	1	4	5	0,094958
		2	7	7	0,403104
		14	692	1600	0,510942
		28	4470	9085	0,534489
		39	804622	1609283	0,549143
		40	8454595	16907773	0,549693
	PL	–	—	-	**0,550000**
Mushroom	IMA_∞	1	23	81	0,007634
		2	29	93	0,013743
		15	590	2866	0,087246
		29	4935	103155	0,097643
		40	19344	712141	0,099198
		44	70142	3945027	0,099636
	PL	-	-	-	**0,100000**
Leukemia	IMA_∞	1	29	69	55,078137
		2	31	71	98,837447
		18	277	731	860,661487
		39	3358	18152	1498,990332
		55	16651	99300	1622,202935
		66	34198	201511	1657,440285
	PL	-	-	-	**1760,587794**
Prostate	IMA_∞	1	84	271	0,137196
		2	86	273	0,883238
		8	186	517	9,514515
		14	544	2118	25,686145
		27	3945	27584	40,587431
		38	17316	124720	45,050108
	PL	-	-	-	**48,315219**
Breast	IMA_∞	1	20	44	0,712290
		2	31	69	6,267710
		17	477	2008	237,448612
		41	4517	25948	321,996845
		61	13845	90521	349,313626
		68	28919	200043	359,589871
	PL	-	-	-	**417,846509**

Tabela II
COMPARAÇÃO ENTRE O IMA_2 E A SOLUÇÃO DO SMV^{light}

Base	Algoritmo	GFMP	Iteracões	Atualizações	Margem
Iris	IMA_2	1	4	5	0,174052
		2	57	156	0,692600
		10	209	512	0,771528
		18	9730	19036	0,789825
		23	10724	20914	0,803810
		28	12252	23970	0,816950
	SVM^{light}	-	-	-	**0,817555**
Mushroom	IMA_2	1	23	81	0,054074
		2	29	93	0,095856
		14	991	4466	0,337148
		18	2101	14655	0,349794
		26	14589	202751	0,361704
		32	88498	2381053	0,364725
	SVM^{light}	-	-	-	**0,365719**
Leukemia	IMA_2	1	29	69	1732,151920
		2	31	71	3151,000863
		9	198	699	10192,507569
		17	988	5291	12514,607498
		29	6600	47316	13054,957828
		33	15221	116196	13152,288305
	SVM^{light}	-	-	-	**13444,174637**
Prostate	IMA_2	1	84	271	6,781741
		2	86	273	43,587395
		8	301	1049	336,283441
		12	904	5157	415,990657
		21	6822	61416	450,010909
		25	13894	135314	456,902969
	SVM^{light}	-	-	-	**472,675499**
Breast	IMA_2	1	20	44	36,711967
		2	32	71	504,269047
		6	79	206	3450,029626
		14	989	5091	4378,271355
		22	13268	76646	4562,511185
		23	27066	158056	4570,891690
	SVM^{light}	-	-	-	**4830,037506**

Tabela III
COMPARAÇÃO ENTRE O IMA_1 E A SOLUÇÃO DE PL

Base	Algoritmo	GFMP	Iteracões	Atualizações	Margem
Iris	IMA_1	1	4	5	0,226000
		2	19	49	0,921692
		21	3097	5540	1,208872
		40	7628	13396	1,281852
		59	281512	556580	1,340415
		65	53850223	107686148	1,349173
	PL	-	-	-	**1,350000**
Mushroom	IMA_1	1	23	81	0,166667
		2	29	93	0,348457
		11	244	678	1,144728
		20	1492	8415	1,439323
		28	37291	565831	1,492092
		31	256626	5409440	1,497753
	PL	-	-	-	**1,500000**
Leukemia	IMA_1	1	29	69	9044,128374
		2	31	71	16988,067861
		49	4242	24483	300035,059822
		85	19745	133826	348873,374194
		118	96205	714627	374765,246972
		135	263248	2014568	383119,195159
	PL	-	-	-	**397844,176325**
Prostate	IMA_1	1	84	271	27,584865
		2	86	273	171,498603
		21	1002	4907	9871,983988
		37	3454	24085	13661,739038
		98	57495	578808	18383,255241
		114	120207	1281945	18870,314772
	PL	-	-	-	**19988,557783**
Breast	IMA_1	1	20	44	283,583978
		2	35	77	12387,526714
		50	1976	7588	139039,397997
		151	19555	87825	190408,269355
		242	199134	995190	215562,974822
		259	499199	2576713	220169,342722
	PL	-	-	-	**230705,132931**

A partir dos resultados, pode-se observar que o IMA$_p$ obteve soluções mais precisas em relação à aproximação escolhida e com menos atualizações, na maioria das execuções, que as obtidas pelo ALMA$_p$. Vale lembrar que o ALMA$_p$ é dependente da escolha do parâmetro α, podendo obter uma aproximação ruim se for escolhido um valor alto e podendo a solução não convergir em um tempo satisfatório se for escolhido um valor baixo. É importante destacar, também, que o IMA$_p$ garante, caso escolhido um valor para α, soluções iguais ou superiores à essa aproximação.

Tabela IV
COMPARAÇÃO ENTRE OS ALGORITMOS ALMA E IMA PARA $p = 2$

Base	α	ALMA$_2$			IMA$_2$		
		It.	At.	Margem	It.	At.	Margem
Iris	0,5	45	118	0,090416	2	2	0,110175
	0,4	80	224	0,099124	2	2	0,110175
	0,3	86	383	0,117973	2	2	0,110175
	0,2	421	1415	0,105981	2	2	0,110175
	0,1	1073	5982	0,116127	182	362	0,121484
Mushroom	0,5	105	1269	0,029715	210	945	0,288894
	0,4	115	1362	0,053297	252	1238	0,304400
	0,3	144	1637	0,085424	315	1901	0,316641
	0,2	172	2004	0,109271	11387	147559	0,359838
	0,1	246	3459	0,160323	615	4134	0,333293
Leukemia	0,5	122	734	0,037511	196	779	0,042208
	0,4	127	1071	0,047563	434	2329	0,054253
	0,3	233	2237	0,049439	190	771	0,047739
	0,2	382	5160	0,055853	219	831	0,051707
	0,1	1412	22927	0,059411	1065	8125	0,058506
Prostate	0,5	547	2486	0,020795	732	4442	0,028114
	0,4	677	3975	0,023868	537	2693	0,023526
	0,3	911	7374	0,027180	579	2979	0,027068
	0,2	1561	17672	0,030083	2535	25372	0,032496
	0,1	4011	74497	0,033049	2531	23445	0,032676
Breast	0,5	117	533	0,045557	157	712	0,057387
	0,4	172	914	0,049812	61	175	0,050208
	0,3	249	1711	0,056568	62	175	0,056934
	0,2	449	4012	0,064492	742	4572	0,068205
	0,1	1677	17317	0,068666	969	6824	0,069442

Tabela V
COMPARAÇÃO ENTRE OS ALGORITMOS ALMA E IMA PARA $p = 6$

Base	α	ALMA$_6$			IMA$_6$		
		It.	At.	Margem	It.	At.	Margem
Iris	0,5	132	484	0,099962	50	144	0,106951
	0,4	320	1209	0,096479	50	144	0,106951
	0,3	551	2741	0,099142	50	144	0,106951
	0,2	1150	7603	0,102231	50	144	0,106951
	0,1	5509	37254	0,104122	50	144	0,106951
Mushroom	0,5	27	952	0,071968	203	987	0,064512
	0,4	42	1932	0,077166	449	6174	0,076260
	0,3	63	4321	0,079440	941	12675	0,078592
	0,2	123	12029	0,081524	727	7216	0,078833
	0,1	334	58474	0,083168	24581	576573	0,083657
Leukemia	0,5	212	1920	0,050929	598	3592	0,050714
	0,4	390	4043	0,053302	355	1840	0,046044
	0,3	787	9162	0,054978	403	2343	0,046684
	0,2	1968	25774	0,055688	1466	10461	0,053894
	0,1	8976	127099	0,056427	1559	10234	0,054309
Prostate	0,5	1594	19887	0,015865	1754	13846	0,014721
	0,4	3098	41326	0,016508	740	5072	0,012541
	0,3	6485	95596	0,016934	3130	26769	0,015527
	0,2	16348	271053	0,017180	1983	17108	0,015343
	0,1	72317	1337308	0,017408	34291	324715	0,017018
Breast	0,5	259	1806	0,053246	1499	10244	0,058606
	0,4	469	3531	0,056311	458	2565	0,052207
	0,3	962	7950	0,059298	1435	9340	0,059132
	0,2	2420	21934	0,060604	1757	11319	0,058430
	0,1	10818	106175	0,061733	2555	16178	0,059851

Tabela VI
COMPARAÇÃO ENTRE OS ALGORITMOS ALMA E IMA PARA $p = 10$

Base	α	ALMA$_{10}$			IMA$_{10}$		
		It.	At.	Margem	It.	At.	Margem
Iris	0,5	312	1130	0,089272	70	201	0,099915
	0,4	536	2374	0,091944	70	201	0,099915
	0,3	936	5429	0,094510	70	201	0,099915
	0,2	2392	15372	0,096543	70	201	0,099915
	0,1	11658	75410	0,098182	70	201	0,099915
Mushroom	0,5	32	1362	0,080589	4067	96230	0,087598
	0,4	54	2870	0,084387	778	8273	0,081963
	0,3	76	6548	0,086458	312	1448	0,069343
	0,2	156	18658	0,087632	7965	241422	0,088596
	0,1	476	91884	0,089098	1584	23073	0,086113
Leukemia	0,5	416	3730	0,049628	334	1552	0,039733
	0,4	769	7876	0,050784	1104	6960	0,048613
	0,3	1650	18477	0,051724	1314	8686	0,049745
	0,2	4399	52744	0,052303	2265	15441	0,052258
	0,1	20638	262021	0,052739	1546	9736	0,050318
Prostate	0,5	3946	50246	0,013445	730	4866	0,009763
	0,4	7687	106398	0,013816	1314	10180	0,011389
	0,3	16514	248418	0,014062	1871	14623	0,011745
	0,2	43709	712592	0,014220	2263	17772	0,012134
	0,1	199385	3550823	0,014330	83282	791682	0,014123
Breast	0,5	484	3347	0,052366	1012	6240	0,049478
	0,4	902	6906	0,054280	4706	31562	0,055894
	0,3	1911	15862	0,055821	522	2965	0,044652
	0,2	5002	44724	0,056942	2103	13946	0,053920
	0,1	23157	219849	0,057586	7290	48105	0,056496

VII. CONCLUSÃO

Neste trabalho foi introduzido um novo algoritmo para a construção de classificadores de larga margem com norma arbitrária. A partir dos testes computacionais, o algoritmo mostrou-se eficiente e robusto e, com isso, foi possível a obtenção de excelentes aproximações das soluções de máxima margem para diferentes valores de norma. Além disso, o algoritmo é baseado no modelo perceptron, o que torna suas compreensão e implementação simples.

REFERÊNCIAS

[1] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

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

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

[4] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–407, 1958.

[5] C. Gentile, "A new approximate maximal margin classification algorithm," *Journal of Machine Learning Research*, vol. 2, pp. 213–242, 2001.

[6] S. Rosset, J. Zhu, and T. Hastie, "Boosting as a regularized path to a maximum margin classifier," *Journal of Machine Learning Research*, vol. 5, pp. 941–973, 2004.

[7] V. Kecman and I. Hadzic, "Support vectors selection by linear programming," in *Proceedings of the International Joint Conference on Neural Networks*, 2000, pp. 193–198.

[8] J. P. Pedroso and N. Murata, "Support vector machines with different norms: motivation, formulations and results," *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1263–1272, 2001.

[9] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[10] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, Cambridge, MA, USA, 1999.