Este artigo é uma evolução do artigo 1, onde foram feitas comparações entre estratégias de maximização de margem utilizando o perceptron.
No Artigo 03, também foi adicionada a SVM, o Classificadores por Grafo de Gabriel e o dataset Sonar, um dataset não linear, que contém 111 padrões de sinais de sonar que descrevem minas e 97 padrões que descrevem rochas. Cada padrão é composto por 60 características.

Artigo 3 - Comparação de Estratégias de Maximização de Margem

Matheus Bitarães de Novaes

I. Introdução

Este artigo tem como objetivo apresentar o modelo do perceptron simples bem como algumas estratégias para a maximização da margem obtida. As estratégias de maximização de margem tem como o objetivo encontrar a superfície de separação que divida o plano da forma mais generalista possível em dados linearmente separáveis. O artigo também apresenta as Support Vector Machines (SVM), e os classificadores por Grafo de Gabriel, métodos que também buscam a maximização da margem de separação.

O trabalho conta com uma revisão da literatura, implementação da SVM, classificadores por Grafo de Gabriel e três algoritmos baseados no perceptron, sendo um deles uma proposição de um modelo de comitê para resolução do problema de maximização de margem. será feita uma comparação estatística do desempenho destes algoritmos de acordo com 5 *datasets*.

II. Revisão de Literatura

A. Perceptron Simples e Estratégias de Maximização de Margem

O Perceptron simples [1] é um classificador linear caracterizado por uma matriz de pesos w, que são multiplicados às entradas X e submetidos a uma função de ativação definida por uma função degrau [2]. Após o treinamento, os pesos w definem a equação de um plano que irá separar os dados das classes do problema.

O processo de ajuste dos pesos w do Perceptron dá-se pela correção dos erros da saída do modelo em comparação com os dados utilizados para o treinamento. Durante o processo os pesos serão corrigidos quando a saída do modelo for diferente da saída esperada, conforme a equação abaixo:

$$w(t+1) = w(t) + \eta e(t) x(t)$$

onde $w(t)$, $e(t)$ e $x(t)$ representam, respectivamente, os valores do vetor de pesos, do erro e do vetor de entrada no instante t [2].

Em problemas linearmente separáveis, podem existir infinitas equações que definam um plano que separe os pontos de diferentes classes. É necessário escolher um plano que separe corretamente os pontos e que esteja posicionado de forma a ser o mais equidistante das duas classe quanto possível. Estratégias de maximização de margem são, portanto, utilizadas para tentar-se encontrar este plano, ou uma aproximação dele, de maneira eficiente. Um artigo que propõe uma abordagem para isso é o de Yoav Freund e Robert E. Schapire [3].

O trabalho propõe o uso de um algoritmo chamado de *voted-perceptron*, que combina o Perceptron com uma variação do algoritmo *leave-one-out* de Helmbold e Warmuth [4]. O algoritmo proposto armazena os vetores gerados durante o treinamento e cria pesos para privilegiar vetores que possuem maior taxa de acerto durante o processo. É realizado um comitê com todos os vetores e cada um vota em uma resposta. Os resultados são ponderados pelo peso de cada vetor. O pseudocódigo pode ser visto na figura 1..

Training

Input: a labeled training set $\langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$
number of epochs T
Output: a list of weighted perceptrons
$\langle(\mathbf{v}_1, c_1), \ldots, (\mathbf{v}_k, c_k)\rangle$

- Initialize: $k := 0$, $\mathbf{v}_1 := \mathbf{0}$, $c_1 := 0$.
- Repeat T times:
  - For $i = 1, \ldots, m$:
    - Compute prediction: $\hat{y} := \text{sign}(\mathbf{v}_k \cdot \mathbf{x}_i)$
    - If $\hat{y} = y$ then $c_k := c_k + 1$.
      else $\mathbf{v}_{k+1} = \mathbf{v}_k + y_i \mathbf{x}_i$;
      $c_{k+1} = 1$;
      $k := k + 1$.

Prediction

Given: the list of weighted perceptrons:
$\langle(\mathbf{v}_1, c_1), \ldots, (\mathbf{v}_k, c_k)\rangle$
an unlabeled instance: $\mathbf{x}$
compute a predicted label $\hat{y}$ as follows:

$$s = \sum_{i=1}^{k} c_i \ \text{sign}(\mathbf{v}_i \cdot \mathbf{x}); \qquad \hat{y} = \text{sign}(s) \ .$$

Fig. 1. Algoritmo *Voted-perceptron*[4]

Outra abordagem para o problema pode ser encontrada no trabalho de Saul Leite e Raul Fonseca[5]. Os autores propõem um conjunto de dois algoritmos chamados *Fixed Margin Perceptron* (FMP) e *Incremental Margin algorithm* (IMA), que busca encontrar a solução para um problema linearmente separável dada uma margem fixa. O algoritmo foi desenvolvido com o intuito de evitar resolução do problema de programação quadrática que é utilizado ao se obter a margem máxima calculada pela SVM. Os pseudocódigos destes algoritmos podem ser vistos nas figuras 2 e 3

Algorithm 1. Primal FMP algorithm.

Input: z_m, $\boldsymbol{w}_{\text{init}}$, γ_{f}, η, T_MAX
$\boldsymbol{w}^0 \leftarrow \boldsymbol{w}_{\text{init}}$, $t \leftarrow 0$
repeat
  for $(i = 1, \ldots, m)$ **do**
    if $(y_i \langle \boldsymbol{x}_i, \boldsymbol{w}^t \rangle < \gamma_{\text{f}} \|\boldsymbol{w}^t\|)$ **then**
      $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t \lambda_t + \eta y_i \boldsymbol{x}_i$
      $t \leftarrow t + 1$
    end if
  end for
until (no mistakes were made) or $(t > T\_MAX)$
return $\boldsymbol{w}^t$

Fig. 2. Algoritmo *Fixed Margin Perceptron* (FMP) [5]

Algorithm 2. Incremental margin algorithm.

Input: z_m, η, δ, T_MAX
$\boldsymbol{w} \leftarrow 0$, $\gamma_{\text{f}} \leftarrow 0$
repeat
  $\boldsymbol{w} \leftarrow \text{FMP}(z_m, \boldsymbol{w}, \gamma_{\text{f}}, \eta, T\_MAX)$
  $\gamma_{\text{f}} \leftarrow \max((\gamma^+(\boldsymbol{w}) + \gamma^-(\boldsymbol{w}))/2, (1+\delta)\gamma_{\text{f}})$
until the convergence of FMP in T_MAX iterations is not achieved
return last feasible $\boldsymbol{w}$

Fig. 3. Algoritmo *Incremental Margin algorithm* (IMA)

Para este trabalho, serão implementadas duas estratégias de maximização de margem. A primeira é o algoritmo *voted-perceptron*, descrito anteriormente e definido em [4]. A segunda estratégia é um comitê de perceptrons, onde o resultado majoritário entre 5 perceptrons definirá a saída do modelo. Espera-se que, com esta estratégia, a saída do modelo seja mais genérica e assertiva do que a avaliação de apenas 1 perceptron.

B. SVM

As SVMs (*Support Vector Machines*) são classificadores com o objetivo de encontrar uma margem de separação capaz de dividir os dados de entrada baseado em suas classes. Para isto, mapeia-se os dados em um plano intermediário, onde os dados são linearmente separáveis e uma otimização é realizada para encontrar um plano que maximize a margem de separação, conforme podemos ver na figura 4. A equação para o calculo da saída de uma SVM é a abaixo:

$$y = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x, Z)$$

[2]

Onde $K(x_i, x, Z)$ é a função de kernel e α_i são os multiplicadores de Lagrange que maximizam a função de otimização abaixo, dados o parâmetro C e a matriz de parâmetros de kernel Z:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y j K(x_i, x_j, Z))$$

[2]

Esta equação depende da definição correta da função de kernel $K(x_i, x, Z)$ e de seus parâmetros. O correto funcionamento de uma SVM depende destas intervenções para obter-se acurácia desejável.

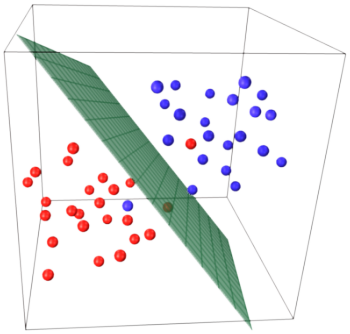

Fig. 4. Hiperplano de separação encontrado após mapeamento das variáveis em uma nova dimensão [6]

C. Modelos de Classificação por Grafo de Gabriel

O Grafo de Gabriel de um set de dados define arestas entre dois pontos p e q se, e somente se, a hiperesfera formada por estes pontos não contiver nenhum outro ponto, como pode ser visto na figura 5.

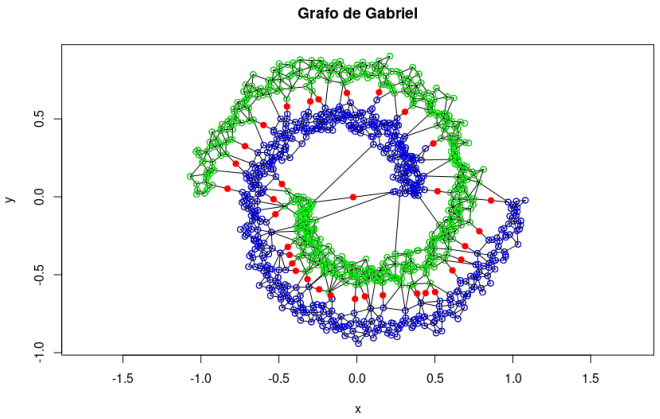

Fig. 5. Ligação entre pontos em um Grafo de Gabriel e margem de classificação definida pelos pontos vermelhos

Esta transformação de dados permite que o *dataset* seja simplificado, removendo os pontos que possuem apenas vizinhos da mesma classe [7]. Ao se transportar o *dataset* para o formato do Grafo de Gabriel, é possível determinar que a margem de separação estará nas arestas dos pontos que possuírem vizinhos de diferentes classes. Esta abordagem permite a simplificação dos dados de treinamento, para posterior aplicação de uma SVM nos dados pré-processados pela transformação realizada

III. Metodologia

Para este trabalho serão utilizados os seguintes grupos de dados:

- *Two Gaussians Dataset*: *Dataset* fictício, artificialmente gerado, com centros em (2,2) e (4,4)

- *Wine Dataset* [8]: *Dataset* com características químicas de vinhos fabricados na mesma região da Itália, porém vindos de três cultivadores diferentes. É uma base de dados com 13 atributos e 3 classes. Para este trabalho, uma das classes foi removida. Desta forma, o problema tornou-se uma classificação binária.
- *Pima Indians Dataset* [9]: Este é um *dataset* com 8 atributos de 768 casos clínicos e duas classes.
- *Iris Dataset* [10]: Este *dataset* contém três classes de 50 instâncias cada, onde cada classe é um tipo de planta iris. Este grupo de dados é composto de 4 atributos que descrevem as dimensões de cada planta. Como este *dataset* possui 3 classes, foi necessário a remoção de uma das classes para que se tornasse um problema de classificação binária.
- *Cervical Cancer Dataset* [11]: Este *dataset* foi coletado no *Hospital Universitario de Caracas* em Caracas, Venezuela e contém informações demográficas, hábitos e histórico medico de 858 pacientes. O *dataset* possui alguns dados vazios que precisam ser tratados.
- *Sonar Dataset* [12]: Este *dataset* contem 111 padrões de sinais de sonar que descrevem minas e 97 padrões que descrevem rockas. Cada padrão é composto por 60 características.

Todos os grupos de dados serão normalizados e os que possuem dados faltantes serão tratados. As linhas que possuem algum campo faltante serão removidas do *dataset*.

Após o pré-tratamento e normalização, os algoritmos serão treinados e avaliados com uma divisão de 70% dos dados para treinamento e 30% dos dados para teste. Este processo será repetido por 50 vezes e os resultados serão comparados.

As comparações entre os *datasets* serão feitas utilizando os testes de Friedman para identificação de diferença estatisticamente significativa em ao menos um dos modelos e, após isto, será realizado o teste de Nemenyi para identificação dos modelos que diferem entre si. O teste de Friedman [13] foi escolhido pois é um teste que não depende que as amostras possuam distribuição normal, que é o nosso caso. Ao se realizar o teste de *shapiro-wilk* para a distribuição de acurácias do primeiro dataset (*iris dataset*), foi identificado que não seguiam a distribuição normal e, portanto, é necessário seguir com um teste não paramétrico, que é o caso do teste de Friedman.

Serão avaliadas as diferenças em valores de acurácia e tempo de treinamento.

IV. Resultados

As três abordagens que utilizam o *Perceptron*, possuem as seguintes configurações:

- Número Máximo de Epocas: 100
- η: 0.01
- tolerância: 0.01

A SVM possui a seguinte configuração:

- $C = 1$
- $\sigma = 1.11$
- Kernel Radial *rbfdot*

A. Duas Gaussianas

Para o *dataset* das duas gaussianas, podemos ver pelo teste estatístico que não há diferença significativa entre as acurácias, porém, nota-se um menor tempo de treinamento para o algoritmo **perceptron**, como pode-se notar pela figura 6

B. Wine Dataset

Para o *Wine Dataset*, nota-se que o teste estatístico indica uma melhor performance para os algoritmos *GG Classification* e *SVM*, que também possuem o maior tempo de treinamento (Figura 7).

C. Pima Indians Dataset

Pela Figura 8 pode-se observar que, para o *Pima Indians Dataset*, os algoritmos *GG Classification* e *SVM* possuem as melhores médias, porém não pode-se afirmar a superioridade na acurácia em comparação com o *perceptron_commitee*. Ja com os demais modelos, a superioridade é estatisticamente significativa.

D. Iris Dataset

Para o *Iris Dataset* pode-se observar pela Figura 9 que não há diferença estatisticamente significativa entre as acurácias. Porém, nota-se que o tempo de treinamento do **perceptron** é inferior aos demais, sendo este algoritmo portanto o indicado para este problema.

E. Cervical Cancer Dataset

Para o *Cervical Cancer Dataset* pode-se observar pela Figura 10 que não há diferença estatisticamente significativa entre as acurácias. Porém, nota-se que o tempo de treinamento do **perceptron** é inferior aos demais, sendo este algoritmo portanto o indicado para este problema.

F. Sonar Dataset

Para o *Sonar Dataset*, nota-se que o teste estatístico indica uma melhor performance para os algoritmos *GG Classification* e *SVM*, que também possuem o maior tempo de treinamento (Figura 7).

Os resultados concatenados podem ser vistos na Tabela I.

V. Discussões

Através dos resultados e testes estatísticos realizados na sessão anterior, pode-se observar que os modelos que buscam maximizar a margem superam o **perceptron** ou possuem resultados equivalentes, apesar de demorarem mais para serem treinados. Para problemas com maior numero de variáveis, a *SVM* e o *GG Classification* obtiveram melhor performance.

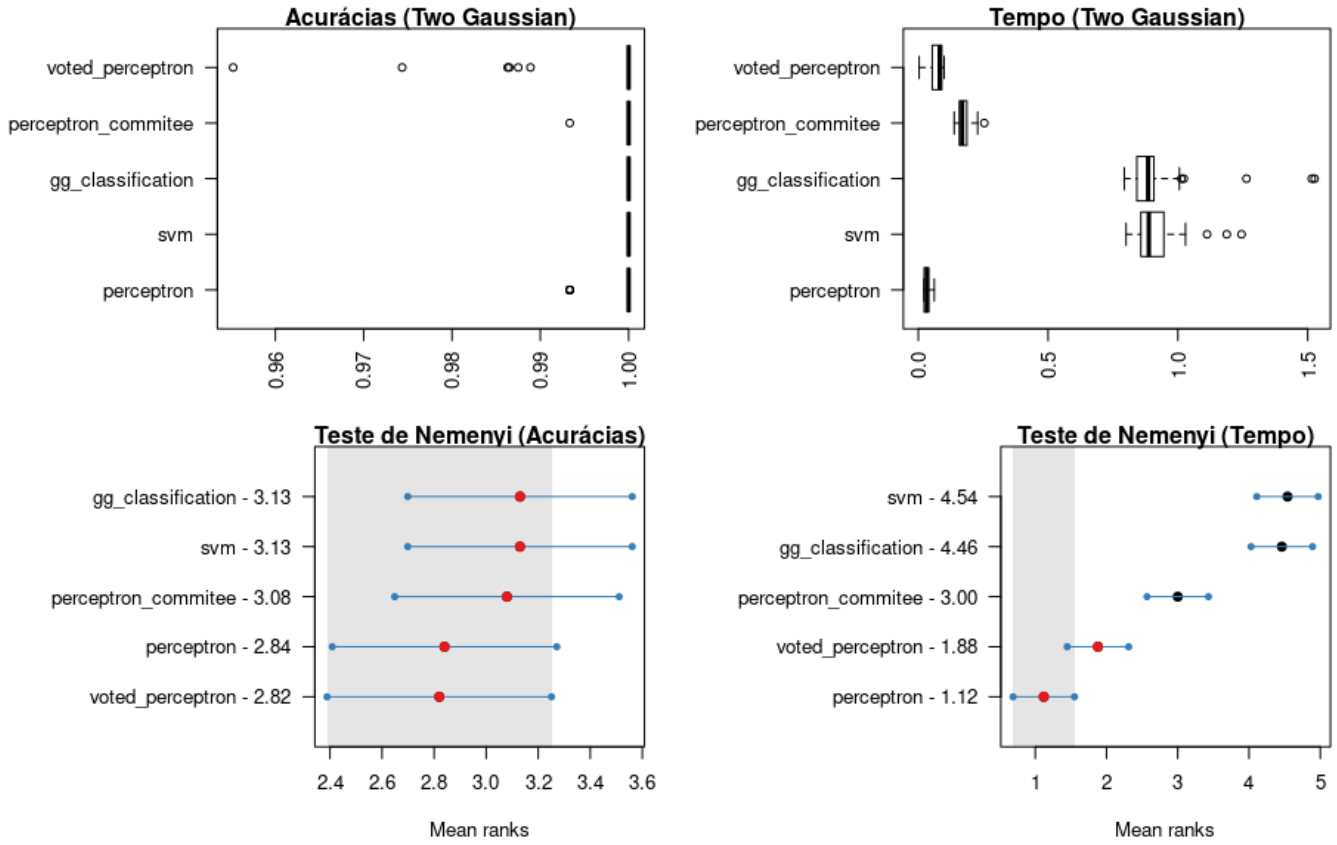

Fig. 6. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *dataset* das gaussianas

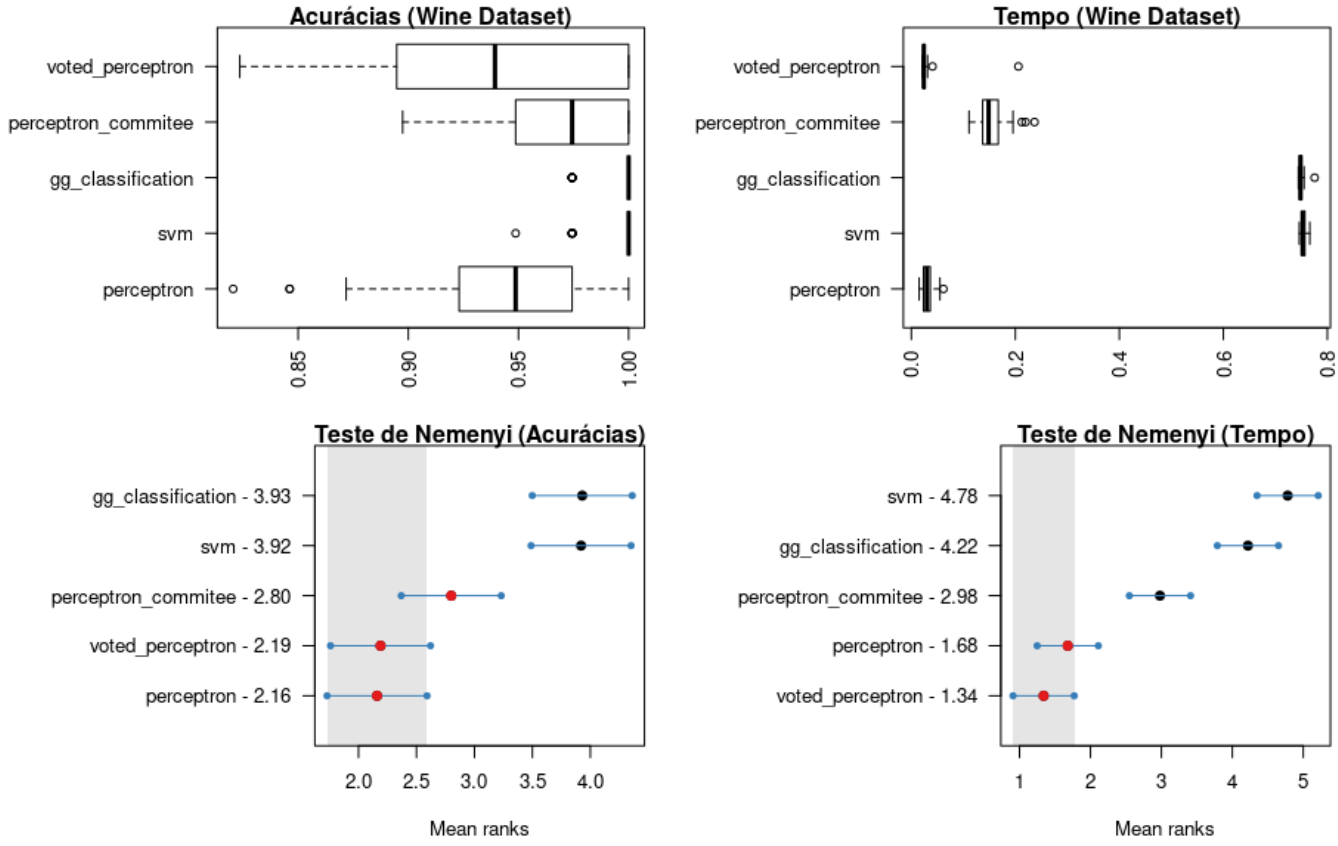

Fig. 7. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *Wine Dataset*

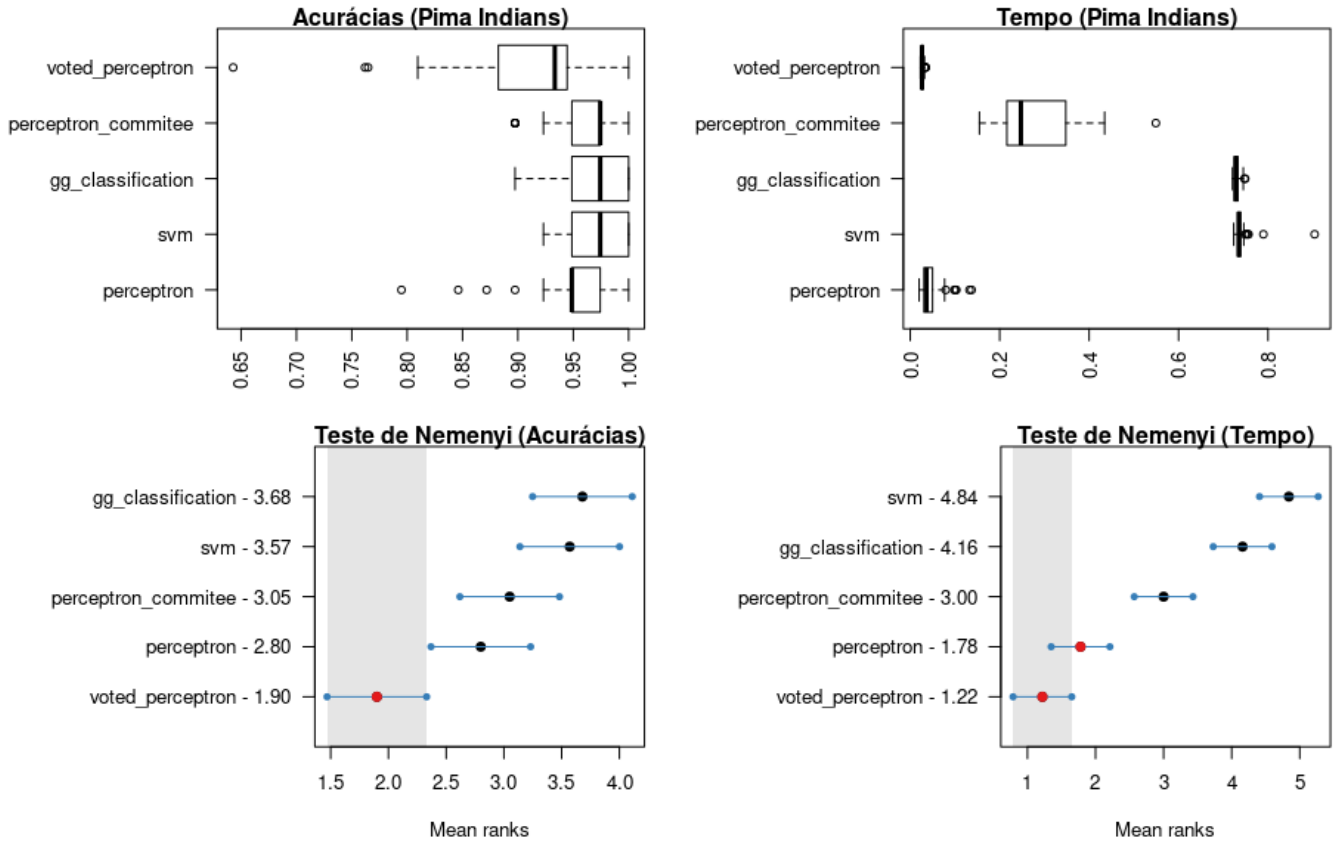

Fig. 8. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *Pima Indians Dataset*

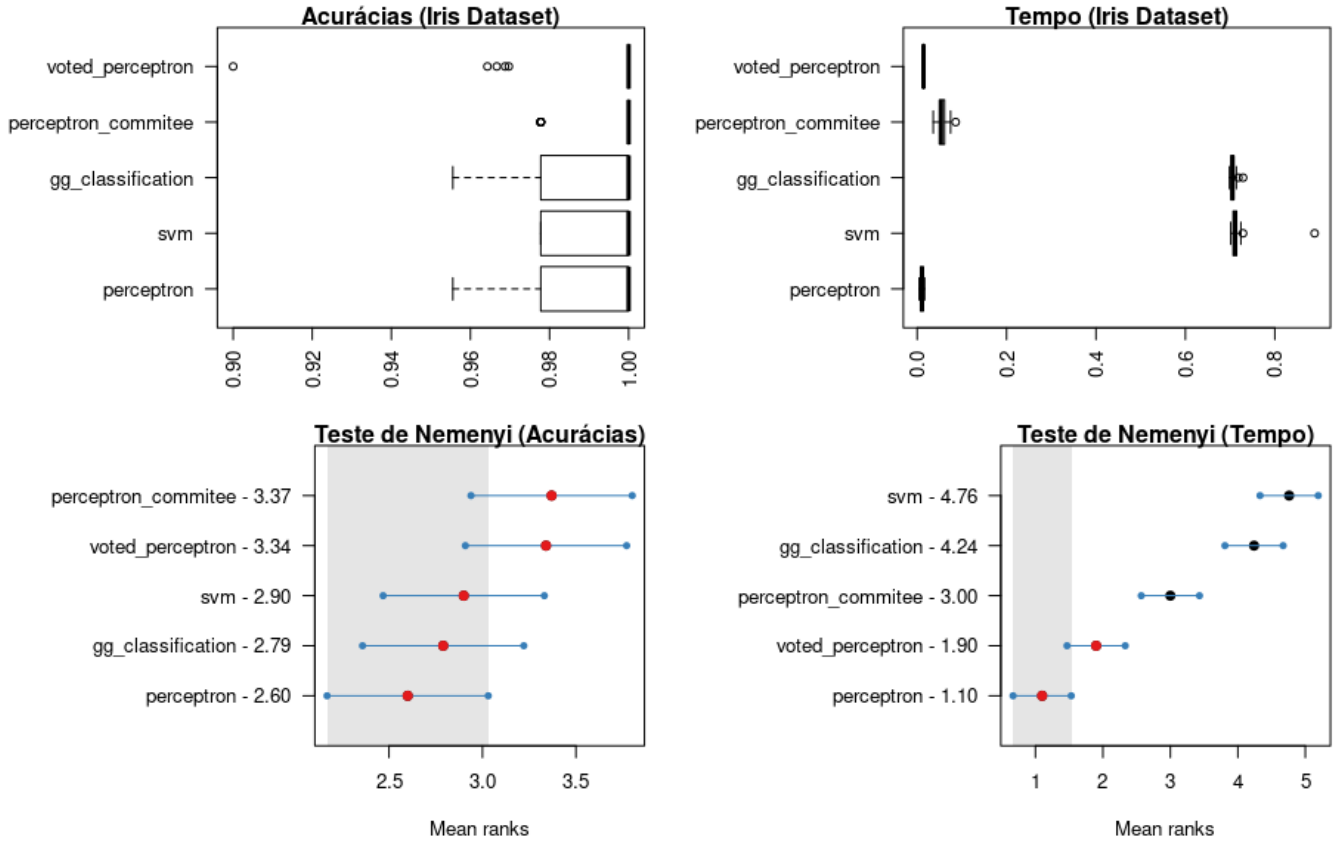

Fig. 9. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *Iris Dataset*

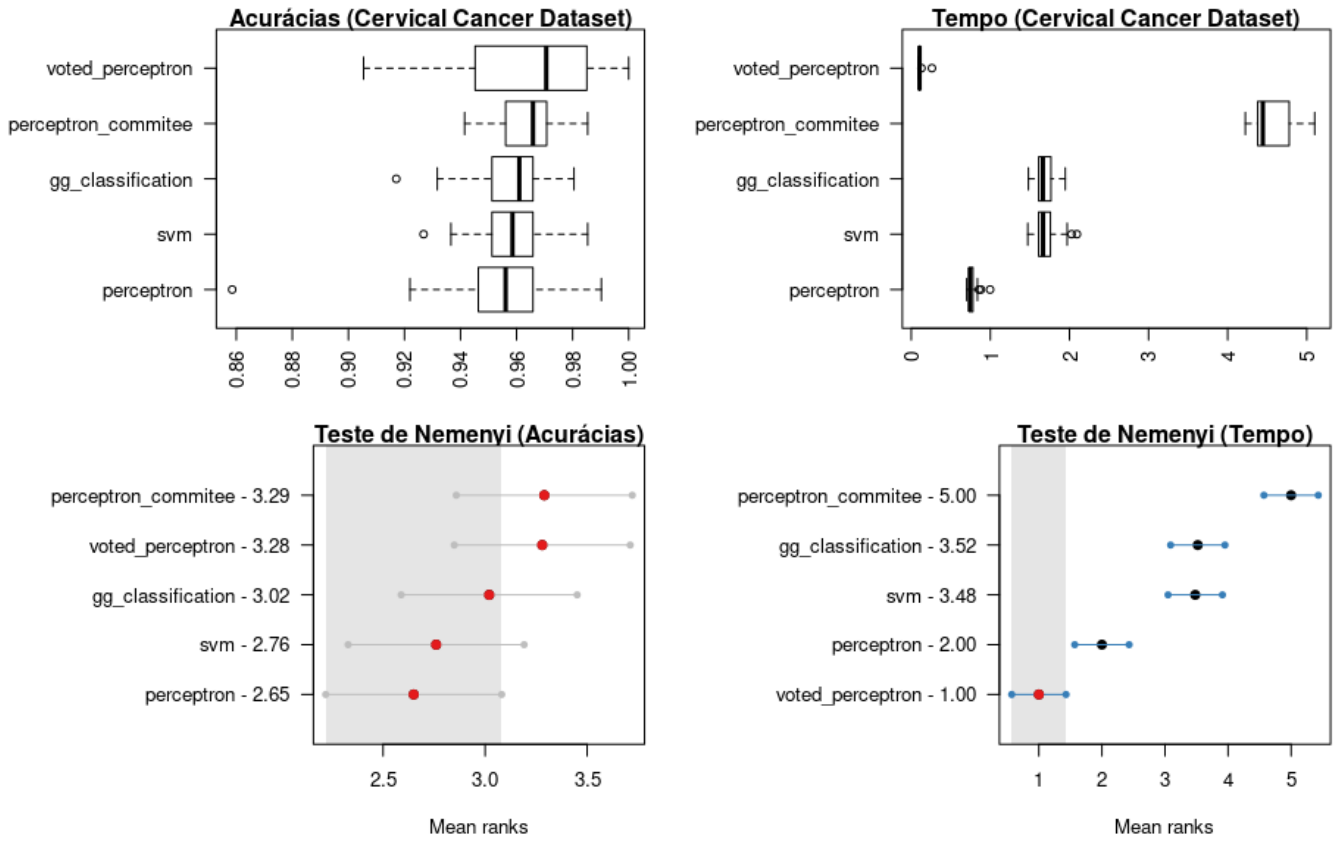

Fig. 10. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *Cervical Cancer Dataset*

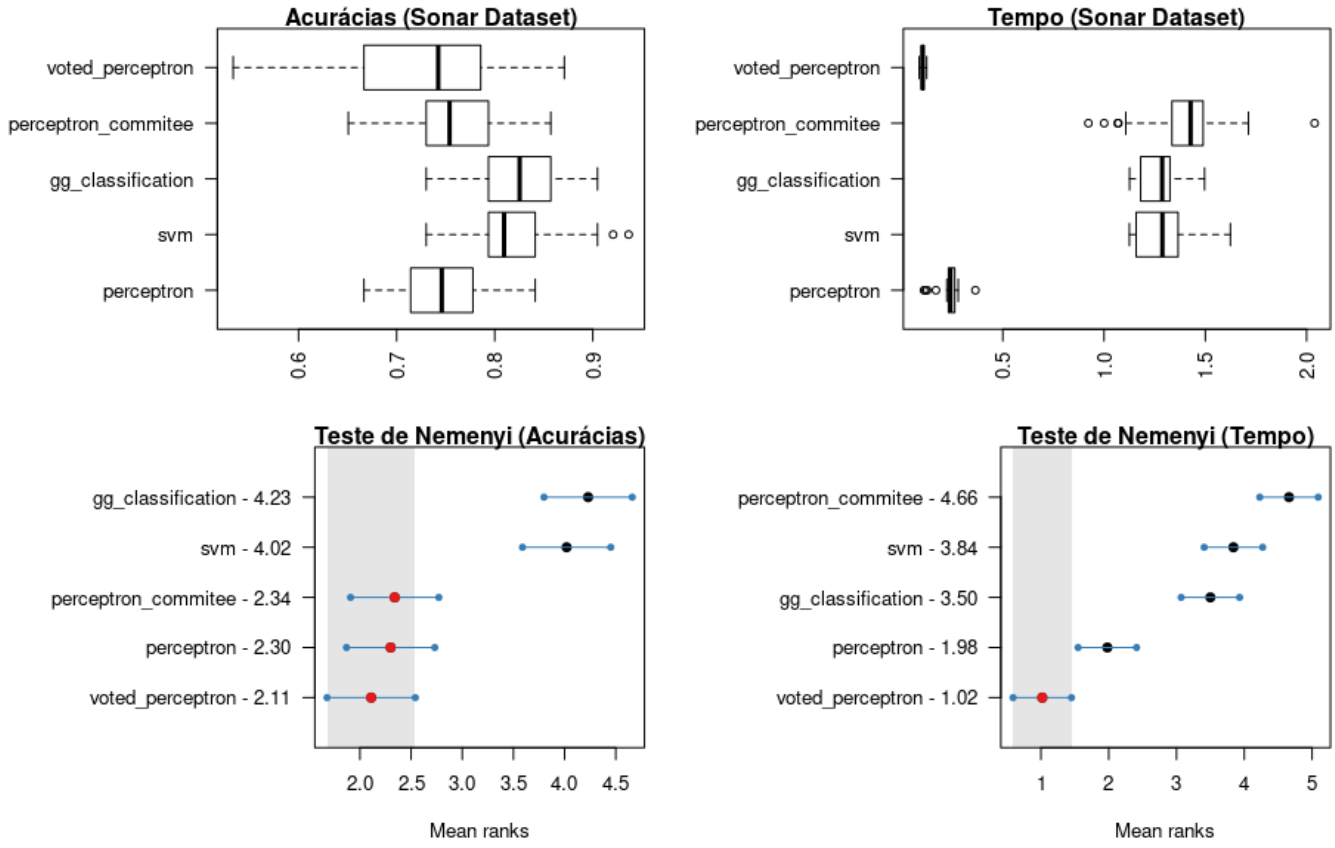

Fig. 11. Boxplot de acurácias, tempos de execução e resultados dos testes estatísticos de Nemenyi para o *Sonar Dataset*

TABLE I
ACURÁCIAS DE ACORDO COM OS MODELOS

Modelo	**Perceptron**	**SVM**	**GG Classification**	**Perceptron Commitee**	**Voted Perceptron**
Duas Gaussianas	0.999 ± 0	1 ± 0	1 ± 0	0.999 ± 0.001	0.999 ± 0.008
Wine	0.945 ± 0.045	0.995 ± 0.012	0.995 ± 0.009	0.970 ± 0.029	0.936 ± 0.054
Pima Indians	0.949 ± 0.037	0.971 ± 0.024	0.969 ± 0.027	0.959 ± 0.032	0.914 ± 0.070
Iris	0.990 ± 0.013	0.993 ± 0.010	0.992 ± 0.012	0.997 ± 0.007	0.995 ± 0.016
Cervical Cancer	0.954 ± 0.021	0.959 ± 0.012	0.958 ± 0.013	0.963 ± 0.010	0.962 ± 0.025
Sonar	0.7479 ± 0.043	0.8210 ± 0.044	0.8222 ± 0.041	0.7546 ± 0.045	0.7331 ± 0.081

VI. CONCLUSÕES

Através deste trabalho pode-se notar a influencia que algoritmos de maximização de margem podem exercer no aumento da acurácia de um modelo, em comparação com o perceptron simples. Foi possível notar também o efeito dessas estratégias no tempo de treinamento do modelo. Os algoritmos de maximização de margem apresentaram acurácias semelhantes e as vezes superiores ao perceptron simples, porém acarretam um maior tempo de treinamento. Portanto, são indicados em contextos onde o aumento da acurácia é mais importante que o aumento no tempo de treinamento

REFERENCES

[1] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[2] A. P. Braga, *Aprendendo com Exemplos: Princípios de Redes Neurais Artificiais e de Reconhecimento de Padrões*. Escola de Engenharia UFMG, 2021, vol. 01.

[3] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.

[4] D. P. Helmbold and M. K. Warmuth, "On weak learning," *Journal of Computer and System Sciences*, vol. 50, no. 3, pp. 551–573, 1995.

[5] S. C. Leite and R. F. Neto, "Incremental margin algorithm for large margin classifiers," *Neurocomputing*, vol. 71, no. 7-9, pp. 1550–1560, 2008.

[6] *Understanding Support Vector Machines: A Primer*, https://appliedmachinelearning.blog/2017/03/09/understanding-support-vector-machines-a-primer/, Acessado em: 2021-07-28.

[7] W. Zhang and I. King, "A study of the relationship between support vector machine and gabriel graph," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 1, 2002, pp. 239–244.

[8] P. Forina M. et al, *UCI machine learning repository - wine dataset*, 1991. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/wine.

[9] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *AAAI*, 2015. [Online]. Available: https://networkrepository.com.

[10] R. Fisher, *UCI machine learning repository - iris dataset*, 1936. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/iris.

[11] A. W. Sobar Rizanda Machmud, *UCI machine learning repository - cervical cancer behavior risk dataset*, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk.

[12] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural networks*, vol. 1, no. 1, pp. 75–89, 1988.

[13] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.