Classificação da doença Ceratocone pelo modelo Support Vector Machine

André Luiz Emidio de Abreu
Centro Universitário Franciscano – FAE
Curitiba, Brasil
ametodosest@gmail.com

Anselmo Chaves Neto
Departamento de Estatística
Universidade Federal do Paraná – UFPR
Curitiba, Brasil
anselmo@ufpr.br

***Resumo*—Este trabalho apresenta uma modificação no processo de ajuste de modelos de *Support Vector Machine* – SVM, sendo aplicados para classificar o grau da doença Ceratocone, que atinge e causa deformidade na córnea humana. O método SVM Correlacionado sugere a introdução de uma pré-fase de ajuste, introduzindo uma constante reguladora baseada no coeficiente de correlação das variáveis K central, KISA, coeficiente de variação, densidade e Hexagonalidade em relação à variável Grau da doença. Os dados foram coletados em exames clínicos e ao todo somam 45 pacientes com nível 1 ou 2 da doença. Os resultados obtidos pelo método SVM Correlacionado apresentaram 100% de acerto nas classificações dos graus dos pacientes, contra 88,89% de acerto obtido pelo método Discriminante Linear de Fisher, demonstrando a eficiência e robustez do método SVM Correlacionado.**

Palavras-chave—Support Vector Machine; coeficiente de correlação; Ceratocone; classificação;

I. INTRODUÇÃO

A córnea possui basicamente cinco camadas, sendo a última delas o endotélio, responsável pela nutrição e metabolismo da córnea. As células dessa camada são em forma de hexágono e com tamanhos parecidos. Para que ele exerça sua função normal, é necessário que as células tenham densidade, tamanho e formas dentro do normal.

O Ceratocone é uma doença degenerativa que atinge a córnea, não inflamatória, bilateral e assimétrica, progressiva, levando às inúmeras alterações na superfície da córnea. Caracteriza-se por afinamento central, protusão apical e astigmatismo irregular, com vários graus de cicatrização, ocasionando uma redução da acuidade visual. A córnea adquire forma cônica devido ao seu afinamento e protusão. Não existe infiltração celular ou vascularização. O cone pode ser de base circular ou elíptica, podendo localizar-se próximo ao eixo visual, superior ou inferior a ele [1].

Para a aplicação do modelo *Support Vector Machine* (SVM) Correlacionado, foram utilizados dados de exames clínicos de 45 pacientes com graus 1 e 2 da doença Ceratocone. Os exames apresentam algumas características como idade e sexo, que não entraram no modelo e os valores das variáveis independentes, que essas sim, fizeram parte da modelagem do SVM: K central, KISA%, Coeficiente de variação, densidade e Hexagonalidade [2][3].

A Fig. 1 apresenta a comparação entre um olho afetado pela Ceratocone (Fig. 1a) e um olho normal (Fig. 1b).

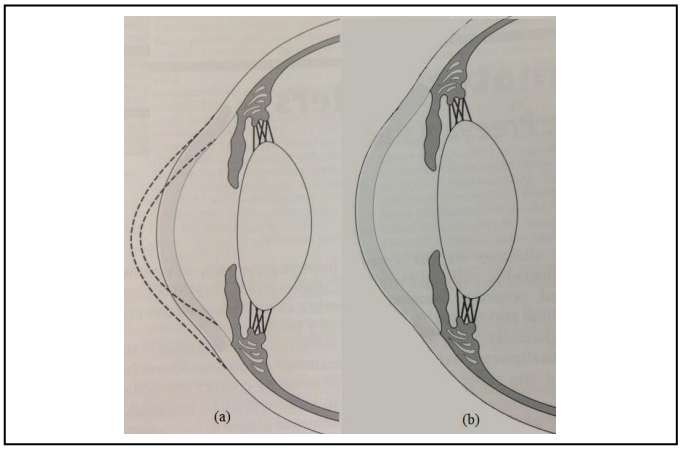

Fig. 1. Olho afetado pela Ceratocone (a) e olho normal (b)

Além da comparação com o método SVM convencional, o método SVM Correlacionado foi comparado com o Discriminante Linear de Fisher (DLF), a fim de verificar sua eficiência perante outros métodos de classificação de padrões.

II. SUPPORT VECTOR MACHINE CORRELACIONADO

O método SVM Correlacionado utiliza todos os pressupostos criados no método SVM convencional, porém, criando uma fase pré-ajuste do modelo, introduzindo o coeficiente de correlação como constante penalizadora aos dados de entrada. A correlação é calculada entre as variáveis colhidas nos exames clínicos e a classe referente ao grau da doença, sendo reservada a classe +1 para o grau 1 de ceratocone e –1 para o grau 2. A solução do modelo SVM foi obtida pela variante *Least Squares Support Vector Machine* (LSSVM) que mantém as mesmas características básicas e a mesma qualidade na solução encontrada que a sua predecessora [4].

Ao contrário do SVM, o LSSVM considera restrições de igualdade no lugar das desigualdades, com isso, resulta um algoritmo que reduz os problemas ao se aplicar a um conjunto extenso de dados [5]. Também, pode-se citar que ao contrário do SVM que utiliza a programação quadrática para calcular seus vetores suporte, que demanda um grande tempo

computacional e a obtenção da solução ótima que possui complexidade considerável, o LSSVM usa um sistema de equações lineares e a função de custo por mínimos quadrados [6][7].

Como para o método SVM tradicional, dado um conjunto de dados de entrada $S = \{(\underline{x}_1,y_1),(\underline{x}_2,y_2),...,(\underline{x}_l,y_l)\} \underline{C} \{X \times Y\}^l$ para o treinamento, ou ajuste do modelo, o objetivo é estimar os parâmetros $\underline{w}$ e b do modelo:

$$y = \underline{w}^t \underline{x} + b, \tag{1}$$

com $\underline{x} \in IR^n$, $y \in IR^n$, $i = 1, 2, ..., n$, $\underline{w}$ são os pesos e b o *bias*.

Logo, formula-se o problema primal:

$$\begin{aligned} \min \quad & \frac{1}{2}\underline{w}^t \underline{w} + \frac{C}{2}\sum_{i=1}^{n} e_i^2 \\ s.a \quad & y_i = \underline{w}^t \underline{x}_i + b + e_i \quad . \\ & i = 1,2,...,n, \end{aligned} \tag{2}$$

com C o parâmetro que penaliza erros altos e é otimizado pelo usuário e e_i são os erros mínimos em relação a reta de separação das classes, conforme verifica-se na Fig. 2.

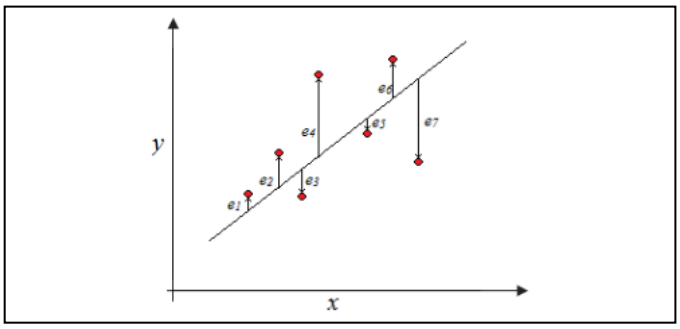

Fig. 2. Exemplo de erros em relação a reta de classificação

Aplicando o Lagrangeano obtém-se o seguinte sistema de equações:

$$\begin{bmatrix} 0 & \underline{1}_n^t \\ \underline{1}_n & (\underline{x}_i \cdot \underline{x}_j) + I_n/C \end{bmatrix} \cdot \begin{bmatrix} b \\ \underline{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \underline{Y} \end{bmatrix}. \tag{3}$$

com $\underline{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)$ e $\alpha_i \in IR$ são os multiplicadores de Lagrange, $\underline{Y} = (y_1, y_2, ..., y_n)^t$ e $\underline{1}_n = (1, 1, ..., 1)^t$.

Para tratar de padrões que não são linearmente separáveis no espaço de entrada utiliza-se a função Kernel trocando o produto escalar, ou interno, $(\underline{x}_i \cdot \underline{x}_j)$ pelo Kernel escolhido. Neste trabalho, utilizou-se o Kernel Gaussiano, ou RBF dado por [8]:

$$K(\underline{x}_i, \underline{x}_j) = e^{-\frac{\|\underline{x}_i - \underline{x}_j\|}{2\sigma^2}}, \tag{4}$$

sendo o valor σ é atribuído pelo usuário, e aplicando ao sistema (3) gera o novo sistema de equações:

$$\begin{bmatrix} 0 & \underline{1}_n^t \\ \underline{1}_n & K(\underline{x}_i, \underline{x}_j) + I_n/C \end{bmatrix} \cdot \begin{bmatrix} b \\ \underline{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \underline{Y} \end{bmatrix}. \tag{5}$$

E a função de classificação é definida por:

$$y = f(\underline{x}) = \sum_{i=1}^{n} \alpha_i K(\underline{x}, \underline{x}_i) + b \tag{6}$$

sendo que se $f(\underline{x}) > 0$ classifica-se o padrão de teste como +1 e para $f(\underline{x}) < 0$, classifica-se o padrão como –1.

Em resumo, o método SVM Correlacionado cria uma pré-fase de treinamento e uma modificação na fase de teste que consiste em calcular a correlação entre as variáveis independentes e as classes dos dados, assim, modifica-se o conjunto de treinamento, destacando as variáveis que possuem uma maior correlação com a variável independente, porém, mantendo-se as variáveis de menor correlação no conjunto de treinamento e teste/validação.

É importante ressaltar que tal modificação não altera o valor absoluto da correlação entre as variáveis. A mudança ocorre caso a correlação seja negativa, que passa a ser positiva. O conjunto de teste/validação também deve ser multiplicado pela correlação.

A correlação entre as variáveis é calculada com base na equação:

$$R = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(y_i^{(p)} - \bar{y}^{(p)}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)}\sqrt{\sum_{i=1}^{n}\left(y_i^{(p)} - \bar{y}^{(p)}\right)}} \tag{7}$$

A Fig. 3 apresenta a sequência do procedimento do desenvolvimento do método SVM Correlacionado para a classificação de padrões.

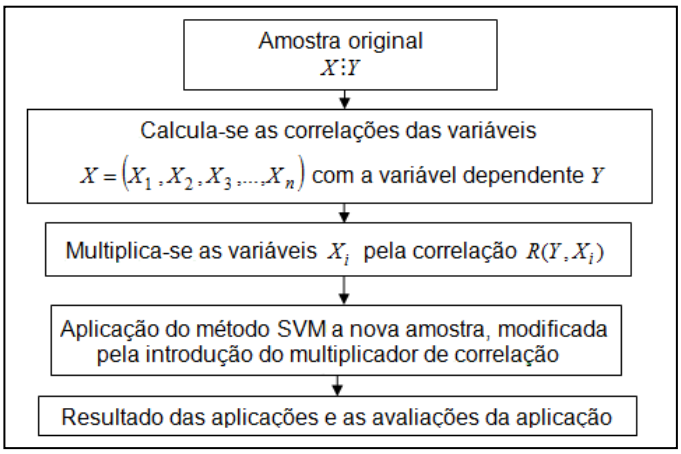

Fig. 3. Procedimento do método SVM Correlacionado

III. Configuração dos Dados Utilizados

A Tabela I apresenta os valores da correlação entre as variáveis K central, KISA%, Coeficiente de variação, densidade e Hexagonalidade e a classe referente ao grau da doença, sendo ela +1 para grau 1 e –1 para grau 2.

TABELA I. Correlações entre as variáveis

	K Central	KISA	CV	Densidade	Hexa.
Classe	-0,8141	-0,3886	-0,3283	-0,2237	0,2185

IV. Discriminante Linear de Fisher

A função discriminante linear de Fisher é uma combinação linear de características originais a qual se caracteriza por produzir separação máxima entre duas populações.

Considerando que $\underline{\mu}_i$ e Σ são parâmetros conhecidos e correspondem respectivamente aos vetores de médias e a matriz de covariâncias comum das populações π_i. Demonstra-se que a função linear do vetor aleatório $\underline{X}$ que produz separação máxima entre duas populações é dada por:

$$D(\underline{X}) = \underline{L}'\underline{X} = [\underline{\mu}_1 - \underline{\mu}_2]'\Sigma^{-1}\underline{X} \quad (8)$$

com $\underline{\pi} = [\pi_1, \pi_2]$, $\underline{L}$ é o vetor discriminante, $\underline{X} = [X_1, ..., X_p]$ vetor aleatório de características das populações $\underline{\mu}$ vetor de médias p-variado e Σ a matriz comum de covariâncias das populações π_1 e π_2.

V. Medidas de Avaliação

Para a avaliação do desempenho do modelo, utilizou-se o coeficiente de correlação *R* e a Raiz do Erro Quadrático Médio (*RMSE – Root Mean Squared Error*). Como mencionado anteriormente, as constantes *C* e σ são determinadas pelo usuário, ou seja, buscam-se os valores que minimizam o erro cometido, assim, uma ampla gama de combinações entre os valores das constantes foram testadas. A combinação considerada ótima deve minimizar o valor do erro *RMSE* e, simultaneamente, maximizar o valor do coeficiente de correlação *R*.

A expressão do erro *RMSE* é dada por [8]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^{(p)}\right)^2} \quad (9)$$

VI. Resultados

Como mencionado nas seções anteriores, o conjunto total de dados, 45 pacientes, foram reservados 27 valores para a fase de ajuste, sendo ela composta por 15 pacientes com grau 1 e 12 com grau 2 da doença Ceratocone. Consequentemente, os exames de 18 pacientes foram utilizados na fase de teste dos modelos, tanto para o modelo SVM Correlacionado, quanto o SVM convencional.

Para a obtenção dos modelos, foi criado um programa computacional em linguagem Fortran 95, onde se testou valores para as constantes *C* e σ, variando os valores de cada uma de 1,0 até 15.000,00, com passo de 0,5 entre uma tentativa e outra, sendo guardada sempre a configuração que apresentava o maior valor para o coeficiente de correlação *R* e o menor erro *RMSE*, e ao final de cada novo ajuste, o valor obtido foi comparado com o valor ótimo até o momento.

Após todos os teste efetuados para os possíveis valores dos parâmetros *C* e σ os valores ótimos para as constantes foram C = 4585,0 e σ = 322,5 acertando 100% dos padrões com a correlação R = 0,8421 e gerando um erro *RMSE* = 0,5730 contra a correlação de R = 0,8035 e erro *RMSE* = 0,6577 do método SVM convencional, com os mesmo valores para os parâmetros *C* e σ, uma vez que também foram os valores ótimos para o modelo convencional.

Vale ressaltar que o método SVM convencional classificou corretamente 100% dos dados, porém, com erro *RMSE* maior e menor valor para a correlação *R*. Sendo assim, é apresentada apenas a configuração assumida como ótima para ambos os casos, uma vez que gerou o menor erro *RMSE*.

A Tabela II apresenta os resultados obtidos pelos métodos SVM Correlacionado e SVM convencional, apresentando as 18 classificações para os pacientes submetidos aos exames, que foram separados para a fase de teste.

TABELA II. Classicações para os 18 Pacientes

Paciente	SVM Correlacionado	SVM	DLF	Classe
1	0,0980	0,0842	+1	+1
2	0,1630	0,1767	+1	+1
3	0,4024	0,4424	-1	+1
4	1,2040	1,0650	+1	+1
5	0,0225	0,0916	+1	+1
6	0,6268	0,2926	+1	+1
7	0,6284	0,3929	+1	+1
8	0,6915	0,6505	+1	+1
9	0,7974	0,6852	+1	+1
10	-0,9295	-0,6741	-1	-1
11	-0,5912	-0,3690	-1	-1
12	-0,2671	-0,2527	-1	-1
13	-0,5666	-0,3585	-1	-1
14	-0,5116	-0,4216	-1	-1
15	-0,4824	-0,5305	+1	-1
16	-1,3532	-1,3878	-1	-1
17	-0,4016	-0,3029	-1	-1
18	-0,1385	-0,0029	-1	-1

VII. Conclusões

Como observado na seção anterior, o método SVM Correlacionado apresentou valores para o erro *RMSE* menores que o método SVM convencional, atém de valores para coeficiente de correlação *R* superiores ao método convencional. Se comparado ao Discriminante Linear de Fisher, o desempenho foi ainda melhor, uma vez que o método DLF

comete o erro ao classificar dois pacientes, um que pertence à classe +1 foi classificado como – 1 e um de classe – 1 foi classificado como +1. Ademais, as soluções dos modelos SVM foram simplificadas utilizando o método LSSVM, onde a complexidade computacional é menor que as soluções de programação quadrática que demandam os métodos SVM.

Assim, verificou-se a eficiência do modelo desenvolvido para este conjunto de dados. Para os próximos trabalhos estuda-se a aplicação do método Correlacionado para problemas de regressão não linear múltipla.

REFERÊNCIAS

[1] R. M. S. Elias; C. Lipener; R. Uras; L. Pavês. "Ceratocone: fatores prognósticos". Arquivos Brasileiros de Oftalmologia. São Paulo, v. 68, nº 4, pág. 491 – 494, Agosto, 2005.

[2] N. A. Feliciano de Deus. "Estudo do endotélio corneal e suas mudanças em pacientes com ceratocone segunto o grau de progressão da ectasia" (Trabalho de conclusão de curso – Optometria). Universidade do Contestado, UnC, Canoinhas, Santa Catarina, 2015.

[3] M. B. Souza. "Rede de aprendizado supervisionado como método de auxílio na detecção do ceratocone" (Tese de Doutorado em Oftalmologia). Faculdade de Medicina da Universidade de São Paulo, São Paulo, 2011.

[4] J. A. K. Suykens; T. V. Gestel; J. D. Brabanter; B. D. Moor; J. Vandewalle. "Least Squares Support Vector Machines". World Scientific Publishing Co. 2002.

[5] L. T. Santos. "Abordagem da máquina de vetor suporte otimizada por evolução diferencial aplicada à previsão de ventos" (Dissertação de Mestrado em Engenharia Elétrica). Universidade Federal do Paraná, 2013.

[6] A. Borin. "Aplicações de Máquinas de vetores de suporte por mínimos quadrados (LS-SVM) na quantificação de parâmetros de qualidade de matrizes lácteas" (Tese de Doutorado em Química). Universidade Estadual de Campinas, 2007.

[7] R. S. Shah. "Least Squares Support Vector Machine". 2005.

[8] A. L. E. Abreu; A. Chaves Neto. "Modelo para previsão de evaporação em reservatórios de água". Anais do Simpósio de Métodos Numéricos Computacionais (SMNC), UFPR, 2015.