Seleção de Parâmetros na Função de Kernel do SVM Utilizando Algoritmo Genético

Autor: Guilherme Castro Silva

RESUMO

O trabalho a seguir trata da seleção de parâmetros na função de kernel do classificador Máquinas de Vetor de Suporte (SVM) utilizando algoritmos genéticos. Para verificar a eficiência do método em relação aos métodos usuais, foi selecionado algumas bases de dados disponíveis na plataforma da UCI Machine Learning Repository. O desenvolvimento do método proposto foi feito utilizando a linguagem R.

***Palavras Chaves* - algoritmos, R, classificação, reconhecimento de padrões, SVM, máquinas de vetor de suporte, algoritmo genético.**

I. INTRODUÇÃO

O presente trabalho busca descrever e comparar a metaheurística desenvolvida para a seleção dos parâmetros (C e γ) na função de kernel do classificador Máquinas de Vetor de Suporte (SVM) utilizando algoritmos genéticos.

II. DEFINIÇÃO DO PROBLEMA

Apesar do algoritmo Máquinas de Vetores Suporte (SVM) apresentar um bom poder de generalização, seu desempenho depende da seleção de parâmetros na função de kernel do classificador. A escolha de parâmetros inadequados pode resultar em decréscimo na acurácia dos resultados.

Atualmente não existe um método universal para guiar a seleção de parâmetros do kernel. Tendo em vista essas dificuldades, faz-se necessário desenvolver uma metaheurística como por exemplo: algoritmo genético, simulated annealing ou Algoritmo Subida da Encosta do inglês Hill Climbing (HC) para a seleção destes parâmetros ao invés de utilizar o *grid search*.

III. MÁQUINAS DE VETOR DE SUPORTE

As Máquinas de Vetores de Suporte (SVMs) constituem uma técnica de aprendizado que vem recebendo crescente atenção da comunidade. Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as Redes Neurais Artificiais (RNAs). Existem diversos exemplos de aplicações de sucesso podem ser encontrados em diversos domínios, como na categorização de textos, na análise de imagens dentre outros.

As SVMs são embasadas pela teoria de aprendizado estatístico, desenvolvida por *Vapnik*. Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

Considere o exemplo da Figura 1. Podemos observar que há vários classificadores lineares possíveis que podem separar as amostras disponíveis sem nenhum erro, mas há somente um que maximiza a margem (maximiza a distância entre o classificador e a amostra mais próxima de cada classe).

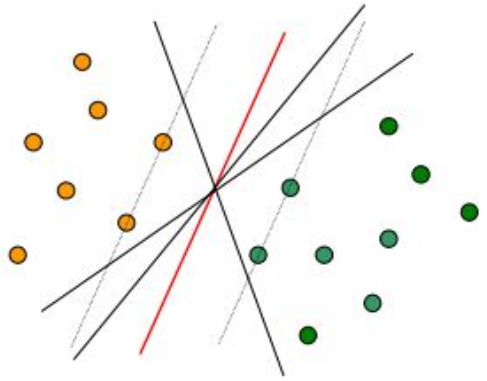

Figura 1: Hiperplano de separação ótimo.

Este classificador linear é chamado hiperplano de separação ótimo (no sentido de apresentar máxima distância entre as classes), pois, ao menos intuitivamente, espera-se que este hiperplano apresenta uma generalização melhor que os demais, quando amostras não utilizadas durante a fase de treinamento devem ser classificadas.

Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classe, ver Figura 1. Para este trabalho, foi utilizado o pacote *e1071* do R que contém a implementação do método *svm.*

IV. ALGORITMOS GENÉTICOS (AG)

Essa classe de algoritmos simulam processos naturais de sobrevivência e reprodução das populações, essenciais em sua evolução. Na natureza, indivíduos de uma mesma população competem entre si, buscando principalmente a sobrevivência, seja através da busca de

recursos como alimento, ou visando a reprodução. Os indivíduos mais aptos terão um maior número de descendentes, ao contrário dos indivíduos menos aptos.

Os requisitos para a implementação de um AG são:

- Representações das possíveis soluções do problema no formato de um código genético;
- População inicial que contenha diversidade suficiente para permitir ao algoritmo combinar características e produzir novas soluções;
- Existência de um método para medir a qualidade de uma solução potencial;
- Um procedimento de combinação de soluções para gerar novos indivíduos na população;
- Um critério de escolha das soluções que permanecerão na população ou que serão retirados desta;
- Um procedimento para introduzir periodicamente alterações em algumas soluções da população. Desse modo mantém-se a diversidade da população e a possibilidade de se produzir soluções inovadoras para serem avaliadas pelo critério de seleção dos mais aptos.

A idéia básica de funcionamento dos algoritmos genéticos é a de tratar as possíveis soluções do problema como "indivíduos" de uma "população", que irá "evoluir" a cada iteração ou "geração". Para isso é necessário construir um modelo de evolução onde os indivíduos sejam soluções de um problema. A execução do algoritmo pode ser resumida nos seguinte diagrama:

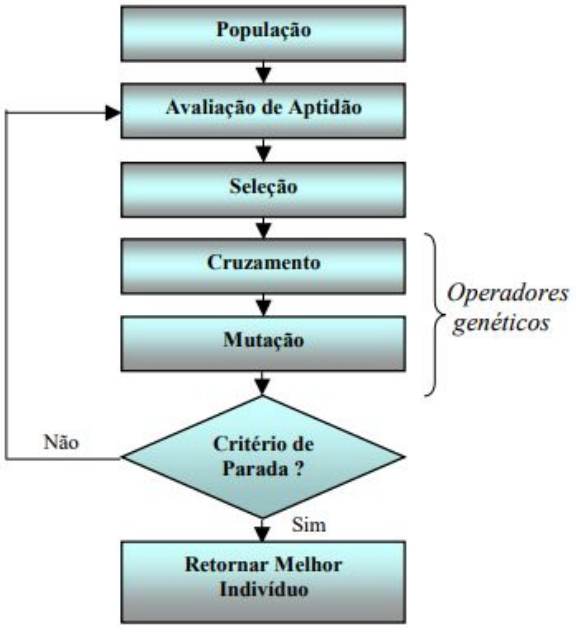

Figura 2: Estrutura básica do Algoritmo.

V. DESCRIÇÃO DO ALGORITMO GENÉTICO

Representação dos Indivíduos

Tendo em vista que o problema consiste em encontrar valores para os parâmetros (C e γ) e verificar qual combinação apresenta a melhor acurácia ao modelo.

O espaço de busca foi estabelecido com seqüências crescentes exponencialmente de C e γ, pois este é um método prático para identificar bons parâmetros. Portanto, foram selecionados 21 valores para "C" e "gamma" 19 valores totalizando 399 combinações distintas como é sugerido na literatura (Tabela 1). Segue abaixo o range de valores escolhido para cada parâmetro.

Parâmetros	C	γ
Valores	$2^{-5}, \ldots, 2^{15}$	$2^{-15}, \ldots, 2^{3}$
Total	21	19

Tabela 1: Número de Parâmetros do Espaço de Busca.

O tipo de representação mais adequada para os indivíduos é a codificação binária onde cada indivíduo, solução candidata, é representada por uma determinada sequência de números binários. Porém, exclusivamente na representação escolhida cada sequência binária conterá apenas um valor '1' e o restante '0'. Segue abaixo a ilustração para cada um dos parâmetros.

0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Figura 3: Representação do parâmetro C.

0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figura 4: Representação do parâmetro γ.

Uma das características da representação binária é que dois pontos vizinhos no espaço de parâmetros não são necessariamente vizinhos no espaço de busca definido pela representação do problema.

Operador de Seleção

Para que seja possível realizar a seleção dos indivíduos, é necessário avaliar a aptidão (*fitness*) de cada indivíduo de modo que a seleção possa ocorrer baseada nas classificações feitas.

Os melhores indivíduos (maior aptidão) são selecionados para gerar filhos através de crossover e mutação.

Operador de Cruzamento

Após a seleção dos indivíduos, é sorteado um valor i tal que pertence ao intervalo [0, 70%] e verifica-se se este valor está dentro da probabilidade de cruzamento, caso esteja, é realizado o cruzamento entre os dois

indivíduos. O objetivo deste operador no processo é propagar as melhores soluções para as futuras gerações.

Após a seleção dos indivíduos, o filho herda com 50% de probabilidade o C do primeiro pai e γ do segundo pai, ou, o filho herda o C do segundo pai e γ do primeiro pai.

Operador de Mutação

Após a recombinação dos indivíduos, é sorteado um valor i tal que pertence ao intervalo [0, 70%] e verifica-se se este valor está dentro da probabilidade de mutação, caso esteja, é realizado a mutação nos dois indivíduos.

O operador de mutação utilizado foi o de encontrar uma nova posição aleatória e atribuir a esta o valor '1' e a posição antiga como '0'.

Critério de Parada

Como não é sabido a solução do problema, podemos definir o critério de parada de duas formas. A primeira seria verificar se a população não está apresentando diversidade em suas últimas X gerações anteriores, isso pode ser verificado a partir do indivíduo médio/ótimo e a segunda maneira que seria definir um número de iterações máximo. O critério adotado foi de verificar se a população não apresenta diversidade em suas últimas 10 iterações e quantidade de iterações máxima igual a 300.

Parâmetros Utilizados

O tamanho da população foi definido como:

20 indivíduos por geração.

O tamanho da população foi definido desta maneira porque a medida que a quantidade de indivíduos aumenta, maior é a quantidade de soluções (possibilidades)

Probabilidade de cruzamento: 70%

Probabilidade de mutação: 70%

VI. TREINAMENTO E TESTE

A base de dados foi dividida em dois conjuntos, um de treinamento e outro para os testes, sendo 70% para treinamento e 30% para teste.

VII. RESULTADOS DE CLASSIFICAÇÃO

O algoritmo foi testado realizando a média da acurácia dentre 30 iterações utilizando cada uma das 15 base de dados, que foram coletadas diretamente do repositório da UCI. Os resultados obtidos perante a metodologia apresentada, serão comparados quando utiliza-se parâmetros aleatórios para a SVM e Grid Search.

	Aleatório	**Grid Search**	**Algoritmo Genético**
Base de Dados	**Acc (%)**	**Acc (%)**	**Acc (%)**
Breast Cancer	0.85668	0.95967	0.96105
Iris	0.90107	0.99892	0.99892
Ionosphere	0.78064	0.93917	0.93870
Liver Disorders	0.63627	0.71669	0.69191
Voting	0.8650352	0.95179	0.9503152
Diabetes	0.7786429	0.93214	0.9466073
Sonar	0.6620584	0.83410	0.8471582
Pima Indians Diabetes	0.6935484	0.75806	0.7664432
Wine	0.8082437	0.98297	0.9874552
Heart	0.6630824	0.84528	0.8034648
Fertility	0.8790323	0.89193	0.8806452
Hepatitis	0.7939646	0.79812	0.8095734
Tic Tac Toe	0.7926747	0.99042	0.9907594
Planning Relax	0.6774194	0.71146	0.7249104
Haberman	0.6795346	0.70650	0.7355896

Tabela 2: Acurácia obtida por cada método dentre as 15 bases de dados.

Através da tabela acima, é possível perceber que o algoritmo genético proposto obteve resultados tão bons quanto o método grid search,+- 3% de diferença , sendo que este último método avalia todas as combinações possíveis.

Além disso, os resultados do algoritmo criado foram muito superiores em relação ao método aleatório, apresentando em alguns casos diferença de 20% de precisão em relação a acurácia.

VIII. CONCLUSÕES

Dentre os métodos utilizados ao longo deste trabalho, ambos os métodos (grid search e algoritmo genético) em conjunto com o SVM apresentaram resultados bastante interessantes. Isto porque o grid

search explora todo o conjunto solução em busca da melhor combinação de gamma e C, ao passo que o algoritmo genético apresenta uma estrutura de vizinhança que permite uma boa capacidade de exploração.

Por último, o método aleatório foi o método menos promissor para este caso. Isso se deve pelo fato de não utilizar algum princípio de validação dos operadores escolhidos.

Pode-se concluir através dos resultados obtidos na Tabela 2 que o modelo de Máquinas de Vetor de Suporte associado ao Algoritmo Genético se mostrou uma técnica bastante eficaz para as bases de dados propostas.

IX. REFERÊNCIAS

[1] Archive.ics.uci.edu. (2017). UCI Machine Learning Repository: [online] Available at: http://archive.ics.uci.edu/ml/index.php [Acessado em 02/07/2019].

[2] R tutorial, http://www.r-tutor.com/r-introduction, [Acessado em 02/07/ 2019].